Fundamentos de Estadística

Silvia N. Pérez

Muchacha en la ventana. Dalí, 1925

Hoja de ruta

- Análisis exploratorio de datos. Variables cualitativas y cuantitativas. Estadísticos de tendencia central y dispersión.
- Nociones de probabilidad. Algunos modelos de probabilidad: distribuciones Binomial, Normal, t, $\chi 2$, F.
- Estimación por intervalos de confianza.
- Pruebas de hipótesis para medias y diferencia de medias. Pruebas de hipótesis para varianza.
- Análisis de la varianza.

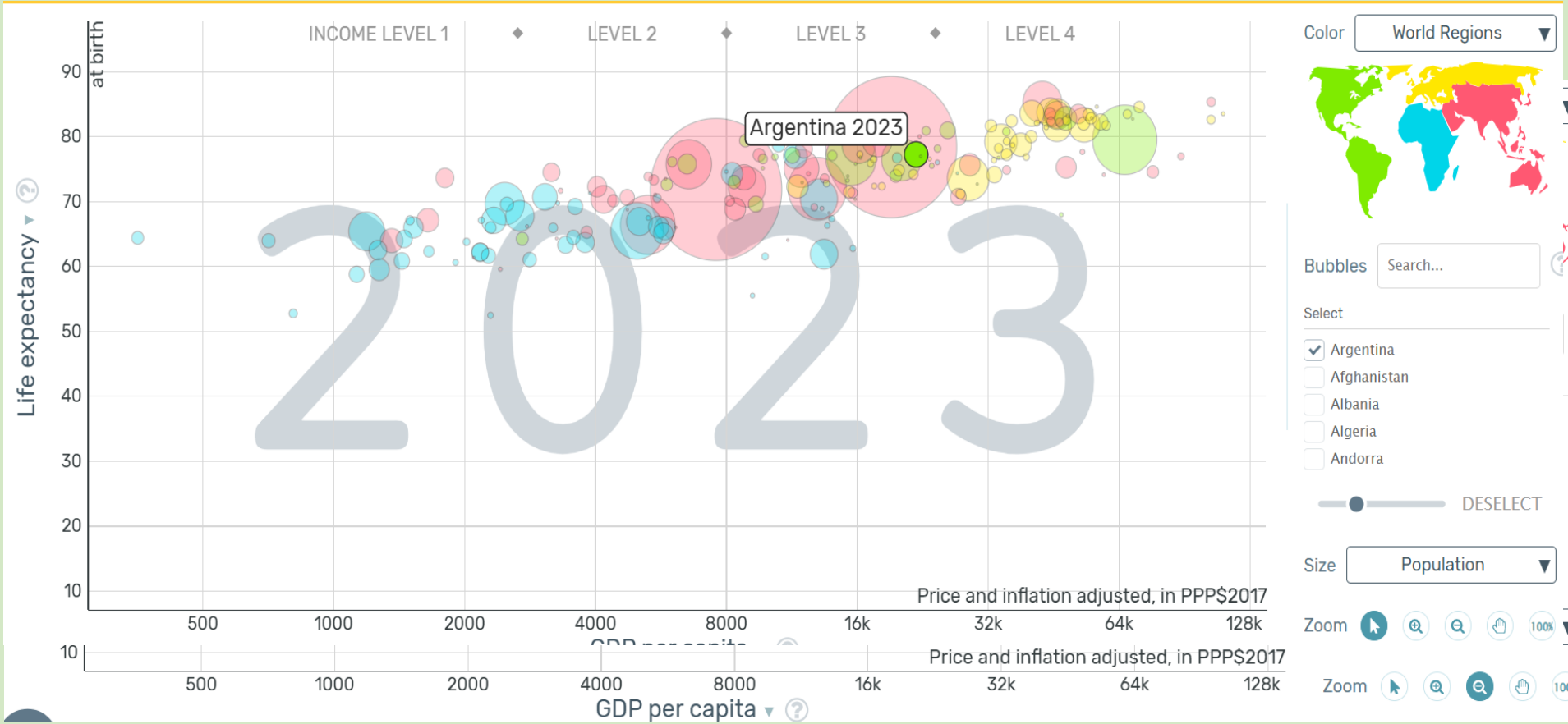
Evaluación

- Trabajo práctico individual: AED + simulación.
- Trabajo práctico grupal: integrador utilizando un conjunto de datos a elección.

¿Por qué necesitamos entender Estadística?

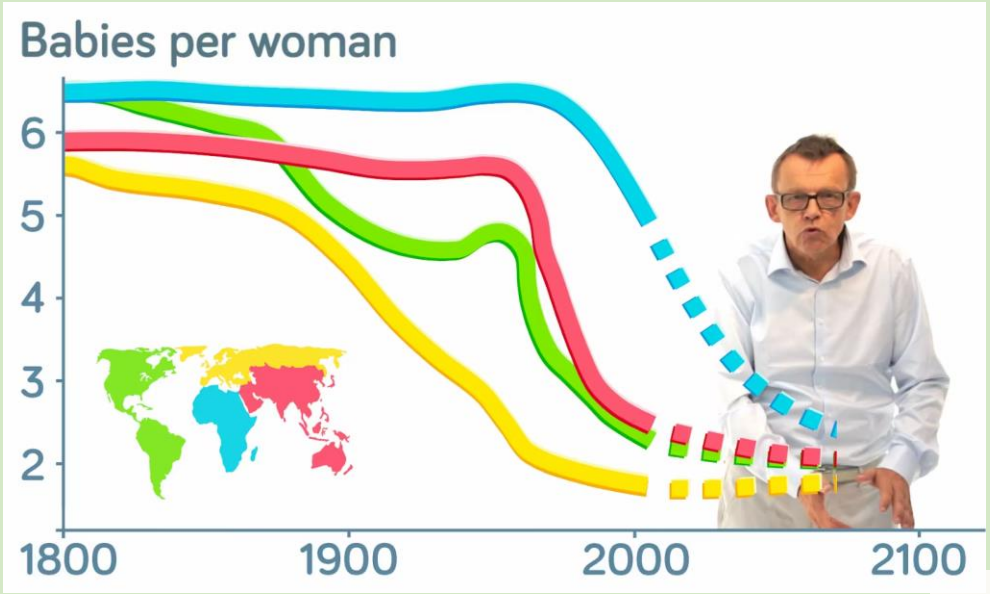
Biblio:
"Probabilidad y Estadística para Ingeniería y Ciencias" Jay L. Devore. Devore%20-%20Septima%20Edicion.pdf
Grolemund, G. y Wickham, H. (2019), R para Ciencia de Datos
https://es.r4ds.hadley.nz/

Para explicar la realidad…

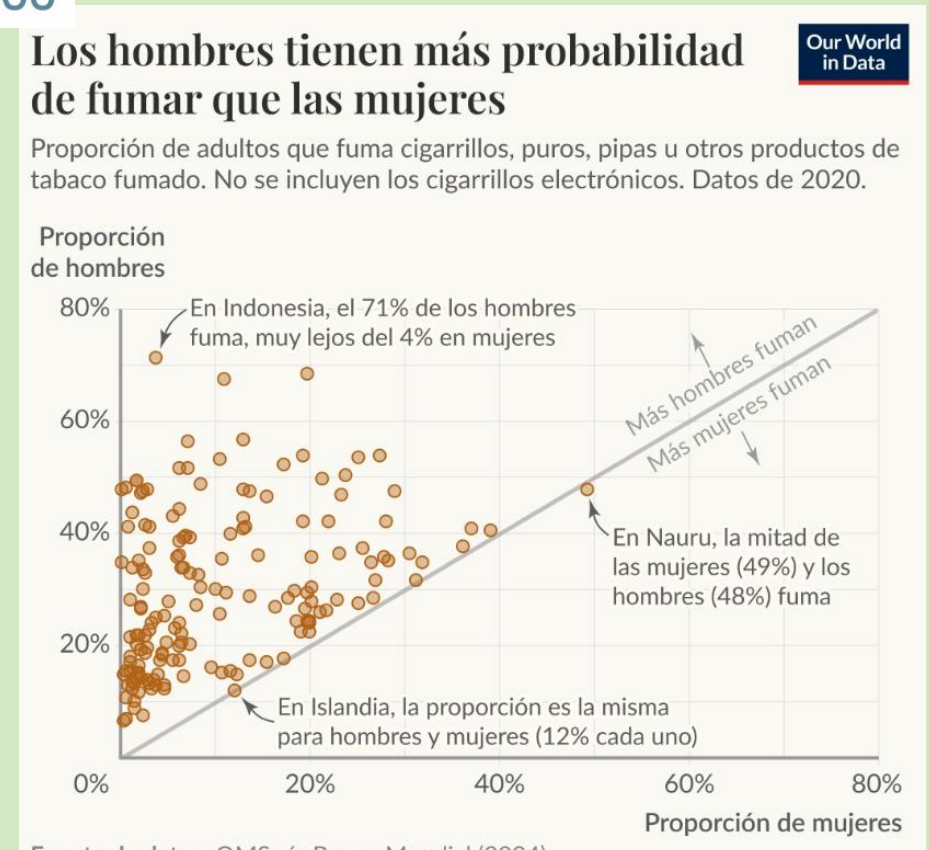


Las burbujas muestran la esperanza de vida y los ingresos medios de todos los países en 2023. Tamaño de la burbuja = #población.

https://www.gapminder.org/tools/#$chart-type=bubbles&url=v2



Para inferir o predecir…



Data.gov

The US Government pledged last year to make all government dat

This site is the first stage and acts as a portal to all sorts of amazi

US Census Bureau

A wealth of information on the lives of US citizens covering popul

Socrata

Interesting place to explore government-related data, with some v

European Union Open Data Portal

As the above, but based on data from European Union institutions

Data.gov.uk

Basamos nuestro trabajo en DATOS

Datos abiertos de la Superintendencia de Seguros de la Nación / Dataset / Recurso

Balances aseguradoras. Recurso completo.

Datos de los balances trimestrales que presentan las aseguradoras a la Superintendencia de Seguros de la Nación (SSN) a través del Sistema de Información de Entidades Supervisadas (SINENSUP).

Campos de este recurso

Título de la columna	Tipo de dato	Descripción
cia_id	Número entero (integer)	Número de registro interno de la entidad asegura
cia_denominacion	Texto (string)	Denominación de la entidad aseguradora o r
indice_tiempo	Fecha ISO-8601 (date)	Trimestre
subramo_id	Texto (string)	Código de subramo.
subramo_descripcion	Texto (string)	Descripción de subramo.
importe	Número entero (integer)	
cuenta_id	Texto (string)	Código de cuenta.
cuenta_descripcion	Texto (string)	Descripción de cuenta.

Buenos Aires > Jefatura de Gabinete > Secretaría de Innovación y Trans...

Datos Abiertos de Buenos Aires

Encontrá todos los datos del Gobierno de la Ciudad en un sólo lugar. Descargalos, analizalos y compartilos.

Buenos Aires Data

A partir del intercambio de ideas y del trabajo colaborativo con profesionales e integrantes de la comunidad de datos, identificamos oportunidades de mejora para elevar la calidad de la información publicada en la plataforma.
Conocé más

¿Qué podés hacer con los datos abiertos?

Descargalos
Desde BA Data podés descargar +400 datasets.

Reutilizalos
Usalos para crear nuevos productos e ideas.

Creá historias
Generá gráficos y visualizaciones
temas que más te interesan de la Ciudad.

Soy Boti,

Cómo obtenemos datos?

Recolección propia, registros, información disponible, recolección de terceros, bases de organismos públicos..... etc!

Algunos sitios:

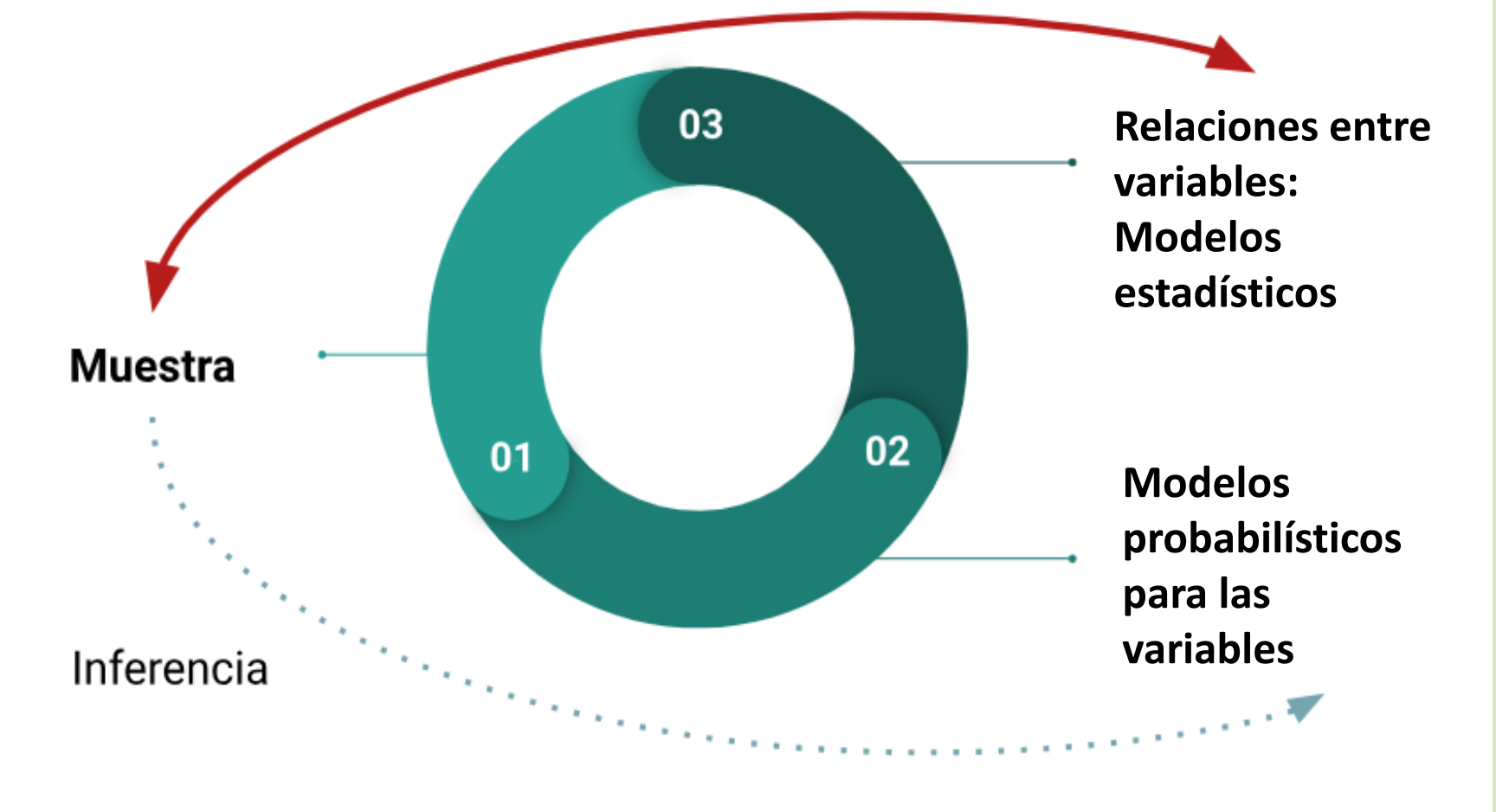
- https://archive.ics.uci.edu/
- https://www.kaggle.com/datasets?fileType=csv
- https://www.datos.gob.ar/
- https://data.buenosaires.gob.ar/
- https://catalogo.datos.gba.gob.ar/

Qué pretendemos de nuestros datos?

Describirlos y explorar → Análisis exploratorio (muestro lo que veo)

Inferir o decidir algo sobre la población de referencia → Inferencia estadística (predigo algo que no veo)

Proceso estadístico



Variables observadas en los datos

Observamos en los datos características o propiedades del objeto de estudio.

Ejemplos:

- Título de grado (Prof. De Matemática, Médico, Lic. en Sistemas, Contador, Ingeniero, …).
- Conocimiento de estadística (nada, poco, mucho)
- Edad.
- Tiempo desde la obtención del título de grado.

Matriz de datos (una forma de verlos)

Casos: Filas

Variables: Columnas

ID	edad	título	residencia	Tiempo dde grado	Estadística?
Juan P.	45	Médico	Lomas de Z	2,4	poco
Manuela G.	52	Contador	CABA	1,7	algo
Laura H	36	Lic. Matemática	Morón	5,5	mucho
..	..	..	..	..	..

Tipos de variables

Cualitativas o Categóricas:

Nominales (Sin orden): Título de grado, Ciudad de residencia, ….

Ordinales (Con orden): Conocimiento de estadística, Posición en el trabajo, …

Cuantitativas o Numéricas :

Discretas: Cantidad de materias aprobadas, Número de hijos, ..

Continuas: Edad, Tiempo de obtención del título de grado, …

Análisis Exploratorio de Datos: univariado

Gráficos

Circulares, barras, histogramas, box-plot, etc.

Tablas

Frecuencias absolutas, relativas, acumuladas, etc.

Medidas resumen

Media, mediana, moda, desvío, mínimo, percentiles, etc.

Gráficos usuales

Univariados

- Sectores o tortas
- Barras
- Histogramas
- Box-plots
- …

Bivariados

- Sectores o tortas combinadas
- Histogramas en paralelo
- Box-plots según categorías
- Dispersión
- Mosaicos
- ….

Representación gráfica (Cualitativas)

Para datos cualitativos se usan:

- Barras
- Sectores

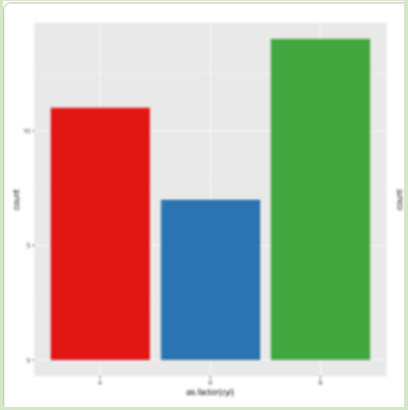


Gráfico de barras:

✓ Se utiliza para variables categóricas y discretas

✓ Existen diversas maneras de realizarlos.

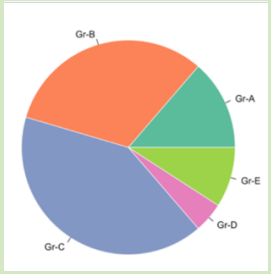


Gráfico circular o sectores:

✓ Se realiza sobre el mismo total.

✓ Puede ser útil en variables nominales u ordinales.

Tablas de Frecuencias p/ variables CUALIS

Frecuencia: Es el número de ocurrencias de un valor o categoría.

Se pueden construir tablas con distintos tipos de frecuencias:

absolutas, relativas , acumuladas (solo tiene sentido si es una variable ordinal)

fre relativa $= f_i/n$

fre acumulada

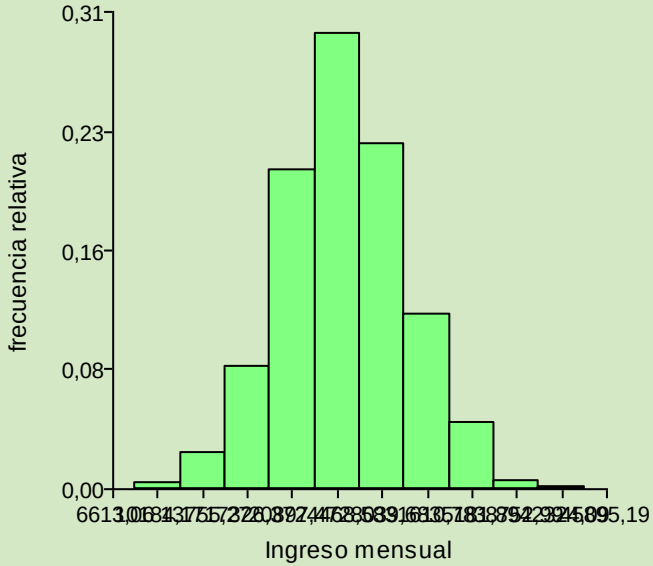
fre relat acumulada

Calificación	f_i	h_i	Porcentaje	F_i	H_i
Excelente	74	0.296	29.6%	74	0.296
Bueno	58	0.232	23.2%	132	0.528
Regular	55	0.220	22%	187	0.748
Malo	63	0.252	25.2%	250	1.000

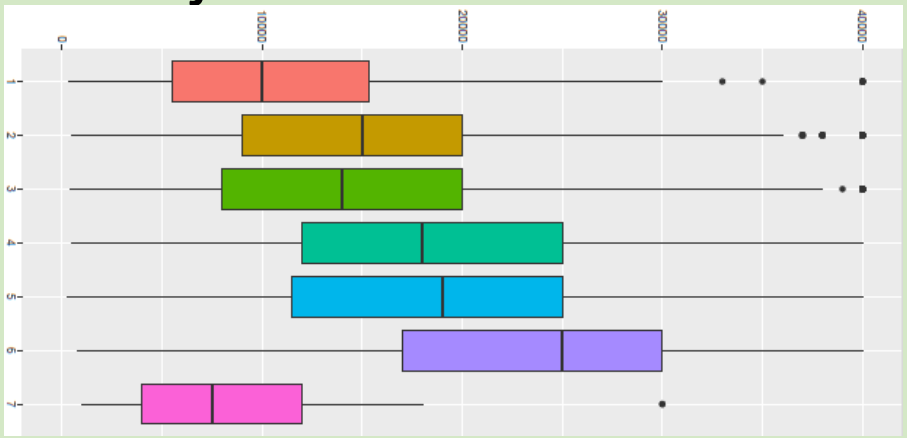
Representación gráfica (Cuantitativas)

Para datos cuantitativos comúnmente se utilizan :

- Histogramas.



- Box-Plots o diagramas de caja.

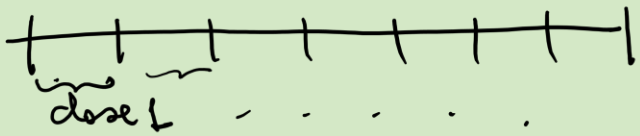


Graficando cuantitativas: histograma

- Se utiliza para variables continuas, o discretas con un amplio dominio.
- Nos permite ver esquemas de comportamientos que son difíciles de ver en una tabla numérica.
- R construye internamente frecuencias (counts) y sobre estas calcula frecuencias relativas que luego reproporciona para graficar valores de ordenadas en cada intervalo que permitan un histograma con área total igual a 1.

Histograma datos de una variable numérica $x_1, x_2, \ldots$

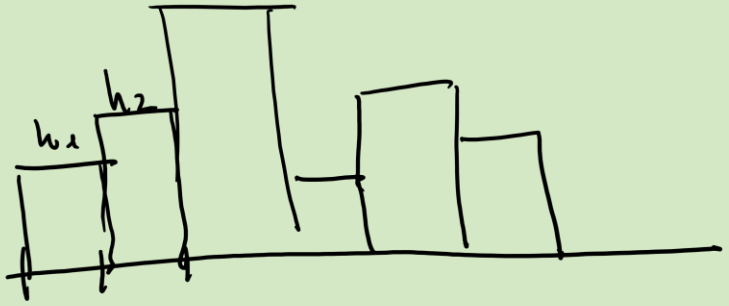
se construyen clases:



$C_i = [a, b)$

amplitud de $C_i = b - a$

clases	frec = f_i	frec rel f_i/n	$\frac{f_i/n}{\text{amplitud } C_i} = h_i$
[,)	3	3/100	
[,)	8	8/100	
	7		
	15		
	⋮		
	100		



se busca q' el área total bajo histograma sea = 1

Ejemplo: “abalone”

Ver datos “abalone.txt”

Ver muestra de análisis básico en AED.html

Más gráficos en:
https://r-charts.com/es/

Medidas resumen

Medidas de tendencia central (posición)

MEDIA (MUESTRAL) o PROMEDIO

cuantis

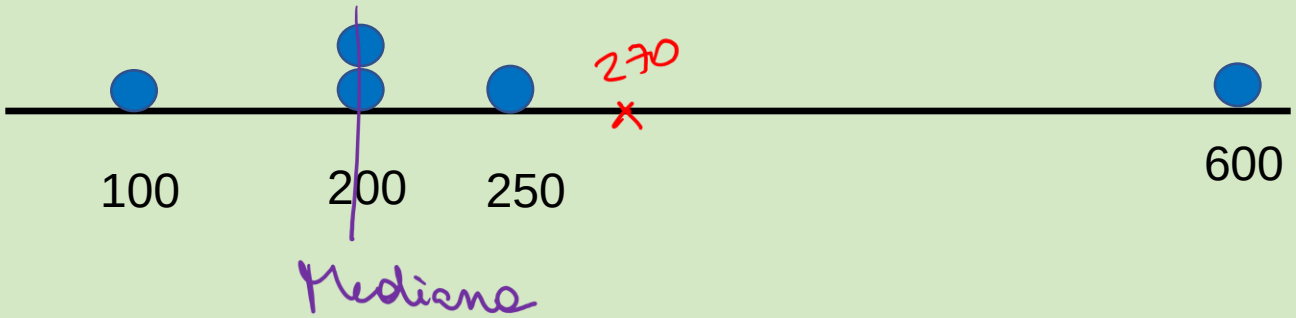
$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

MEDIANA: valor que separa la muestra en dos partes iguales

MODA: valor que se repite con mayor frecuencia en los datos. → usualmente se usa p/ cualis

Medidas de posición

Ejemplo (de juguete): ventas diarias en miles de $ de un kiosco



$\overline{X}$ = Media= (100+2*200+250+600)/5 = 270

(promedio)

Mediana= 200

100, 200, 200, 250, 600

Moda = 200

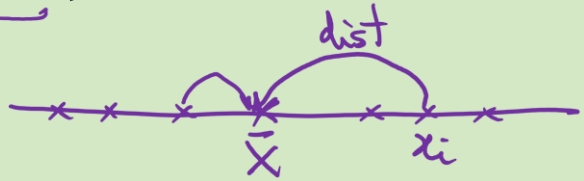
Obs: la media se deja influir p/ datos alejados o atípicos

Medidas de dispersión

VARIANZA (MUESTRAL): calcula un promedio corregido de las distancias al cuadrado respecto del promedio.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Xi - \bar{X})^2$$

dist



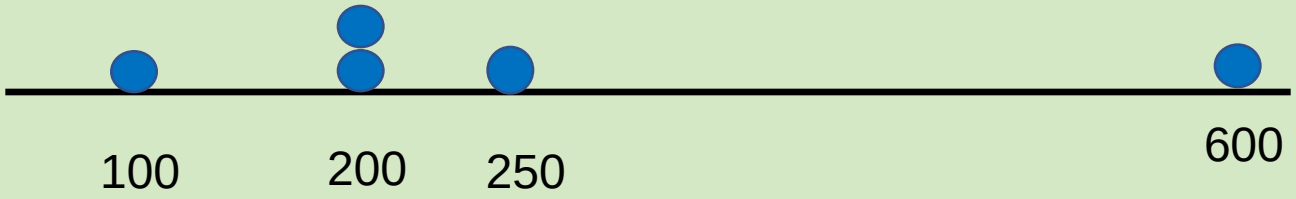
Desvío estándar:

$$S = \sqrt{S^2}$$

Rango: valor máx –valor mínimo

Medidas de dispersión

En el ejemplo



Rango = 600-100 = 500

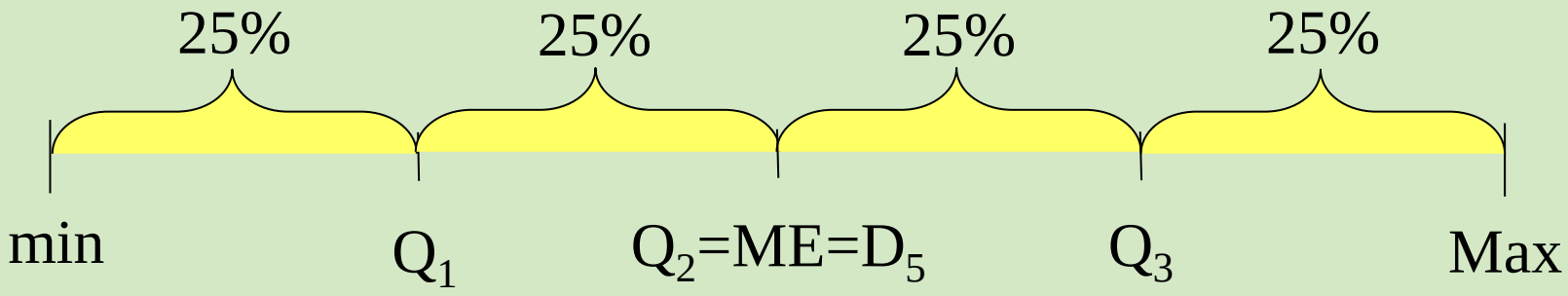
Varianza = la varianza muestral s^2 = 37000 $= \frac{1}{4}\left[(100-270)^2 + (200-270)^2 + \dots\right]$

Desvío = 192,35

Coef de variación = desvío/media= 0,7124

Otras medidas

- Cuartiles: separan a la muestra en cuatro partes iguales.



- Quintiles: separan a la muestra en 5 partes iguales.

- Deciles: separan a la muestra en 10 partes iguales.

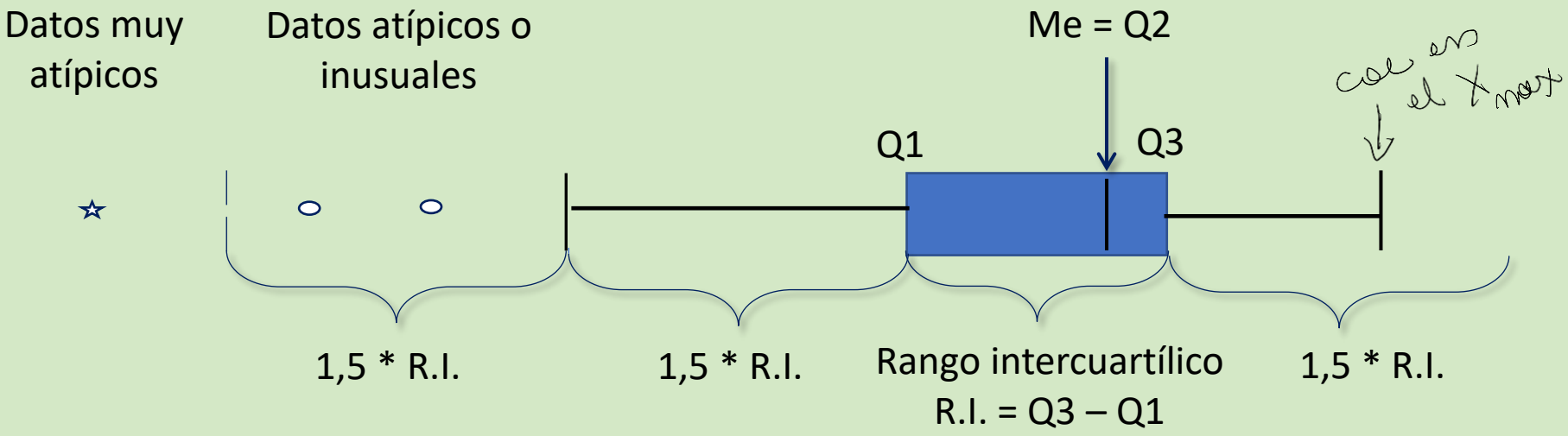
El Cuartil 2 coincide con la Mediana y con el decil 5.

Percentil-k: Es el valor de la variable que deja k% de los datos por debajo de este. Equivale a cuantiles, medido en %.

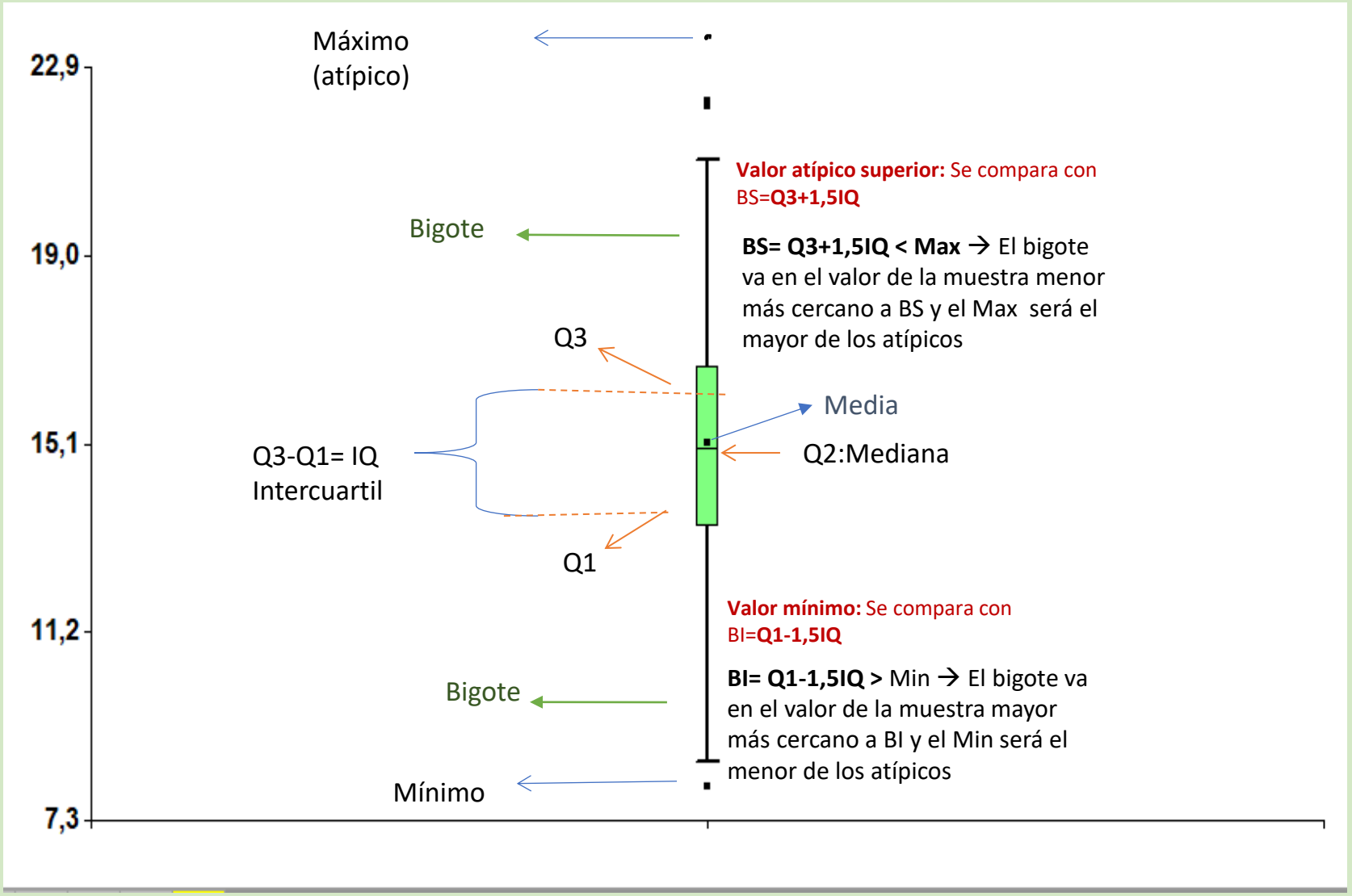
Ejemplo: Percentil 30, P_{30} equivale al cuantil 0,3

Boxplot (diagrama caja-bigote)

- Relacionado con las medidas de resumen.
- Es útil para detectar datos atípicos .
- Se puede observar la asimetría de la distribución de una variable.
- En el gráfico quedan determinadas medidas de posición.

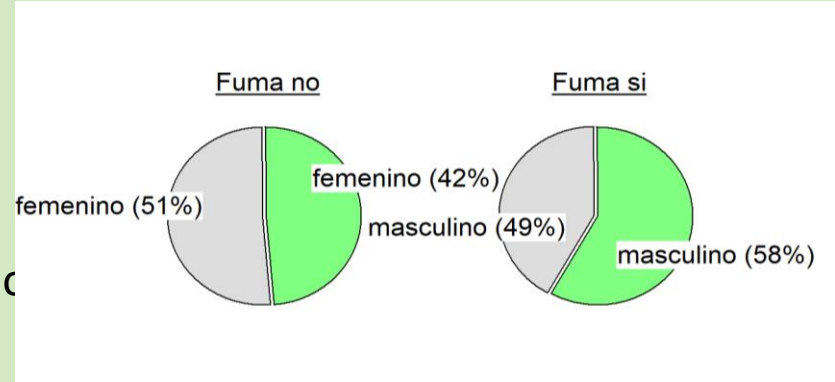


Boxplot

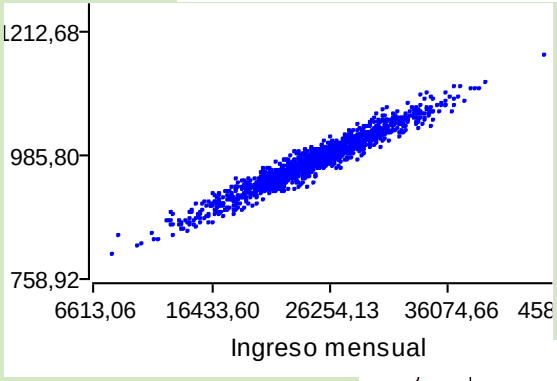

Máximo
(atípico)
22,9
19,0
15,1
11,2
7,3
Valor atípico superior: Se compara con BS=Q3+1,5IQ
Bigote
BS= Q3+1,5IQ < Max → El bigote va en el valor de la muestra menor más cercano a BS y el Max será el mayor de los atípicos
Q3
Media
Q3-Q1= IQ
Intercuartil
Q2:Mediana
Q1
Valor mínimo: Se compara con BI=Q1-1,5IQ
Bigote
BI= Q1-1,5IQ > Min → El bigote va en el valor de la muestra mayor más cercano a BI y el Min será el menor de los atípicos
Mínimo

¿Cómo graficar conjuntamente dos variables?

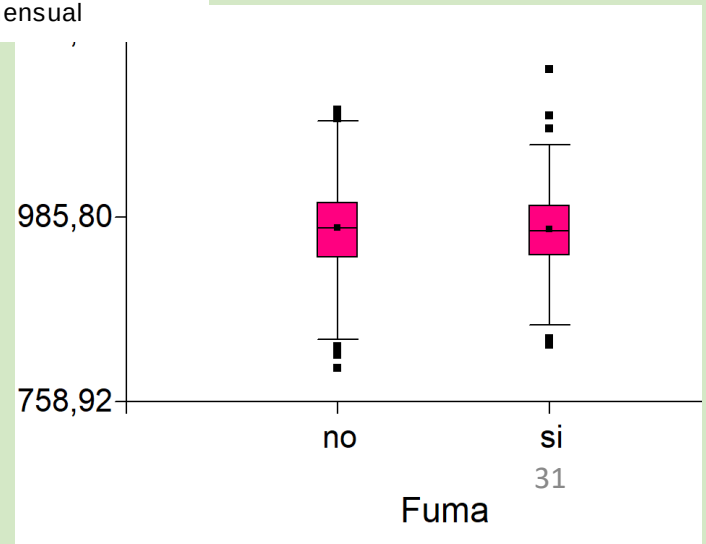
- Si las dos son cualitativas:
  - tablas cruzadas
  - sectores separando por categorías/particio
  - mosaicos



- Si las dos son cuantitativas:
  - Gráficos de dispersión



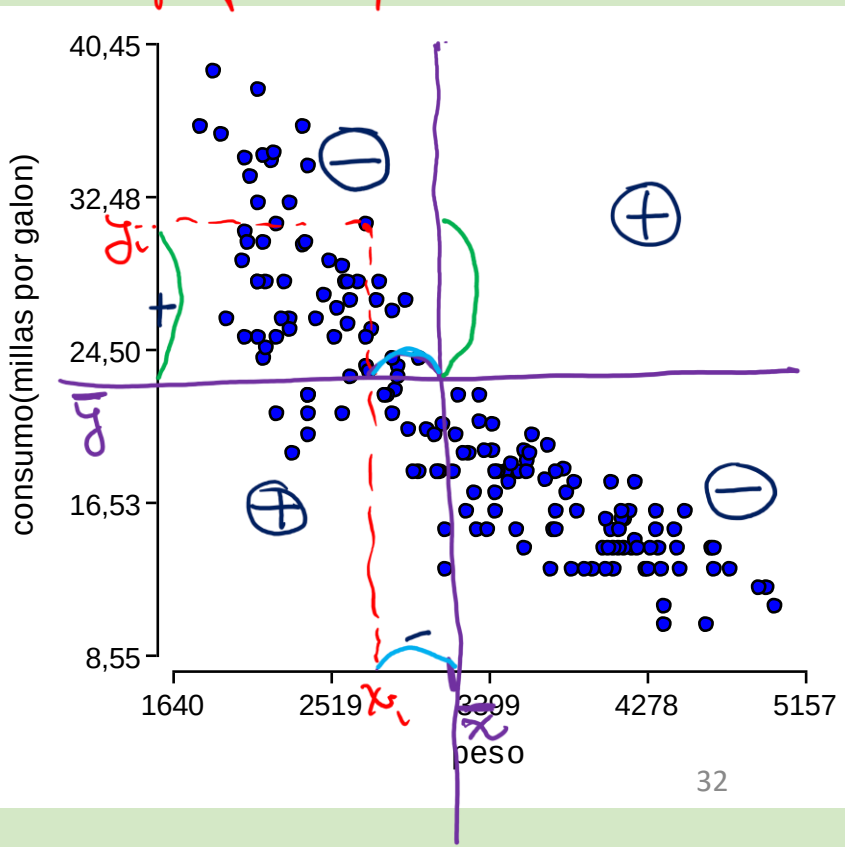
- Cuali - cuanti:
  - Histogramas separando por categorías/particiones
  - Box-plots separando por categorías/particiones



Medida de asociación entre dos cuantis: **Coeficiente de correlación**

Mide el grado de asociación (lineal).

$$r = \frac{\sum_i \left[(x_i - \bar{x}) \, (y_i - \bar{y}) \right]}{\sqrt{\sum_i (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$



Prop
$-1 \le r \le 1$

en el eje → r es (−)

Continuemos con el ejemplo: “abalone”

Ejemplo 1: “premios”

- Identificar variables/tipos
- Gráficos univariados
- Gráficos bivariados
- Tablas!
- Medidas resumen

IMPORTANTE!

¿Preguntas de investigación???

Ejemplo 2: “Futbolistas”

- Identificar variables/tipos
- Gráficos univariados
- Gráficos bivariados
- Tablas!
- Medidas resumen

IMPORTANTE!

¿Preguntas de investigación???