ESTADÍSTICA BÁSICA

Data BootCamp

Cámara Comercio Bilbaio

Germán Alonso Lascurain
germanalonso@opendeusto.es

CÁMARABILBAO C2B campus to business ai aipower tigloo

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS

ESTADÍSTICA DESCRIPTIVA. INTRO

- ¿Qué es un dato?
  - Un dato lo es cuando:
    1. Tiene una característica (variable).
    2. Pertenece a un individuo (registro).

- Los datos no tienen porqué ser números, pueden ser etiquetas (color pelo, estado civil, ...)

ESTADÍSTICA DESCRIPTIVA. INTRO

- Tipos de datos:

  - **Datos cualitativos:** No se pueden medir numéricamente (Color ojos, estado civil, ...). En esencia NO son números, son etiquetas.
    - Nominales: Son etiquetas y entre ellas no existe un orden natural o razonable. Este tipo de datos se pueden clasificar, PERO NO ORDENAR (de mayor a menor, p.ej). (Estado civil, raza, profesión, dirección, ...).
    - Ordinales: Estos sí pueden clasificarse y ordenarse siguiendo una lógica razonable y objetiva (nivel estudios (primaria, secundaria, fp, ...)).

ESTADÍSTICA DESCRIPTIVA. INTRO

- Tipos de datos:

  - **Datos cuantitativos:** Son en esencia un número. Pueden ser continuos o discretos. Se pueden clasificar, ordenar Y **MEDIR DISTANCIAS** entre ellos. La clasificación para variables continuas (altura por ejemplo) se suele hacer por intervalos, pero para las discretas no sería necesario (# hijos).

ESTADÍSTICA DESCRIPTIVA. INTRO

Estadísticos que podemos aplicar a los datos

- Datos cualitativos Nominales:
  - Frecuencia absoluta (cuántos hay).
  - Frecuencia relativa.
  - Moda.
  - NO CALCULAR LA MEDIA, NI VARIANZA, ya que no tiene sentido aplicar estos estadísticos a una etiqueta.
  - La representación gráfica básica son los **diagramas de barras** de frecuencias absolutas o relativas y **gráfico de sectores** (no se recomienda utilizar, ya que no son tan fáciles de interpretar como los de barras. Nuestro cerebro lee mejor las longitudes que las áreas).

ESTADÍSTICA DESCRIPTIVA. INTRO

Estadísticos que podemos aplicar a los datos

	Frec. Abs	Frec. Abs. Acum
P	15	15
S	20	35
G	10	45
M	10	55
D	5	60

- <u>Datos cualitativos Ordinales:</u>
  - Frecuencias absolutas.
  - Frecuencias relativas.
  - Frecuencias acumuladas.
  - Medianas. La mediana (valor central para un conjunto de datos) para individuos impares es un valor, pero para individuos pares son 2 valores, ojo.
  - Moda.
  - NO CALCULAR LA MEDIA, NI VARIANZA, ya que no tiene sentido aplicar estos estadísticos a una etiqueta.

ESTADÍSTICA DESCRIPTIVA. INTRO

Estadísticos que podemos aplicar a los datos

- Datos Cuantitativos:
  - Los estadísticos previos.
  - Media. Es el centro de gravedad de los datos. Tiene el problema de que es sensible a las observaciones extremas (se deja llevar por ellas). La mediana, por el contrario no se deja llevar por estas observaciones y en ocasiones es mejor para representar el valor central de una distribución.
  - Rango (amplitud). Valor máximo – Valor mínimo de los datos. Es muy sensible a los valores extremos (outliers).

ESTADÍSTICA DESCRIPTIVA. INTRO

Estadísticos que podemos aplicar a los datos

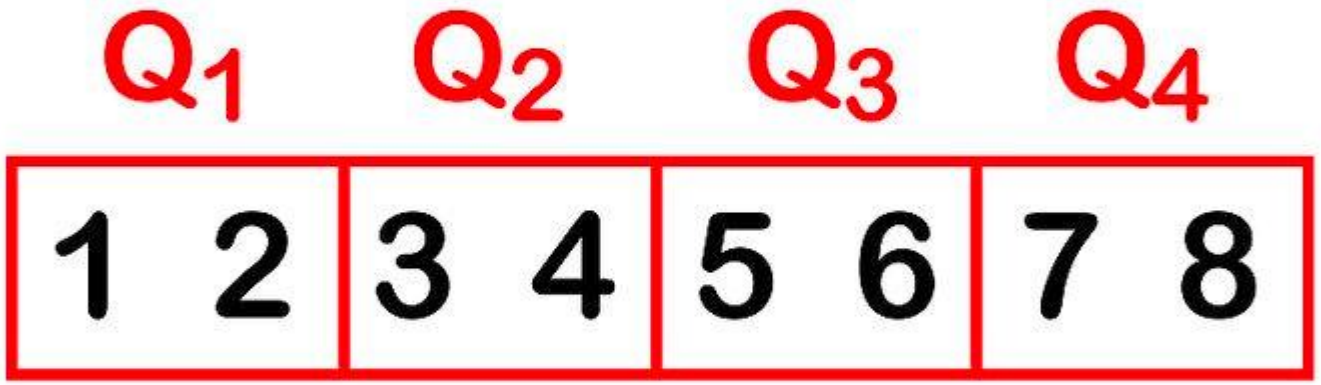
- Datos Cuantitativos:

  Para eliminar outliers, se usa el rango intercuartílico que es rango de datos entre el Primer y Tercer cuartil.

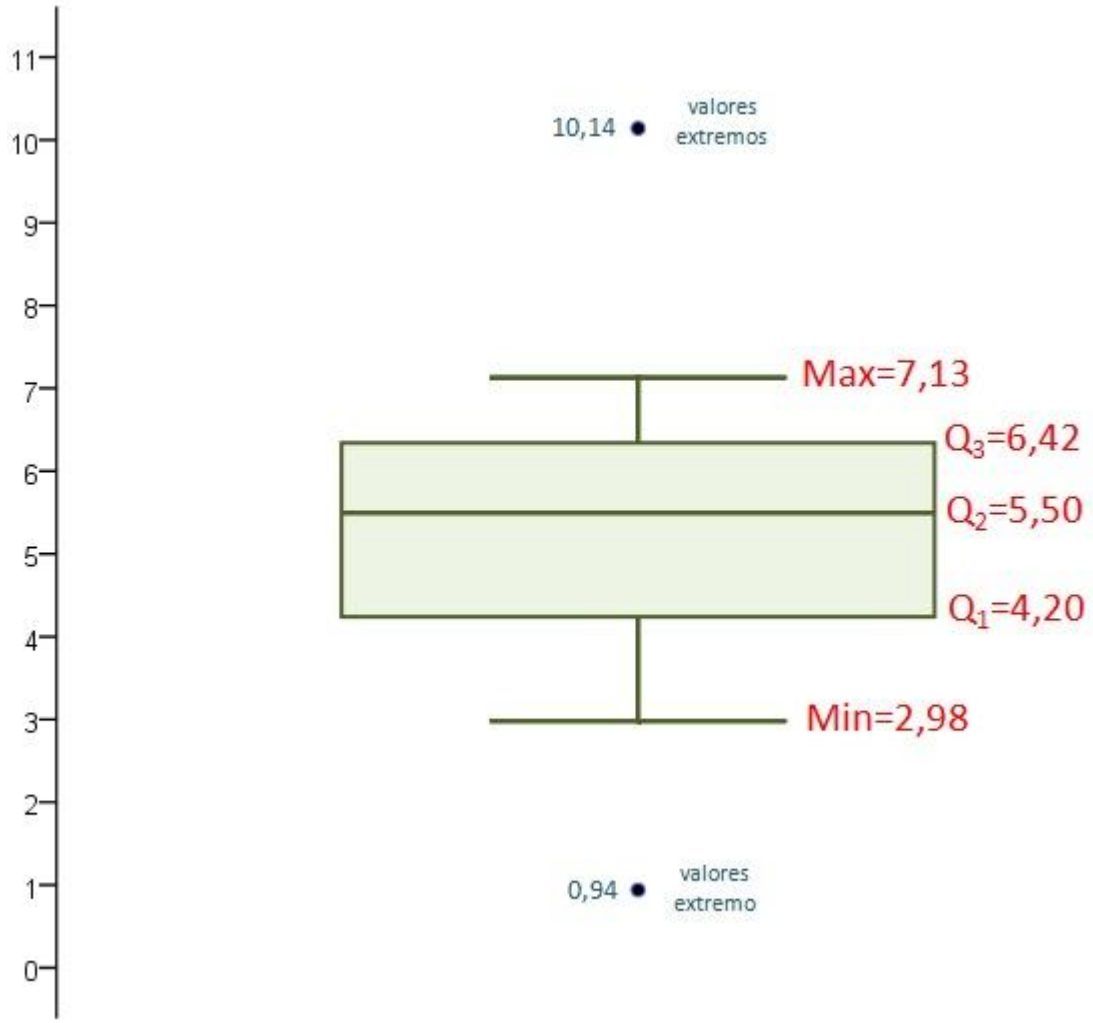
  $Q_3 - Q_1$ es el rango intercuartílico.

Los cuatiles, nos permiten conocer rápidamente la dispersión y la tendencia central de un conjunto de datos.

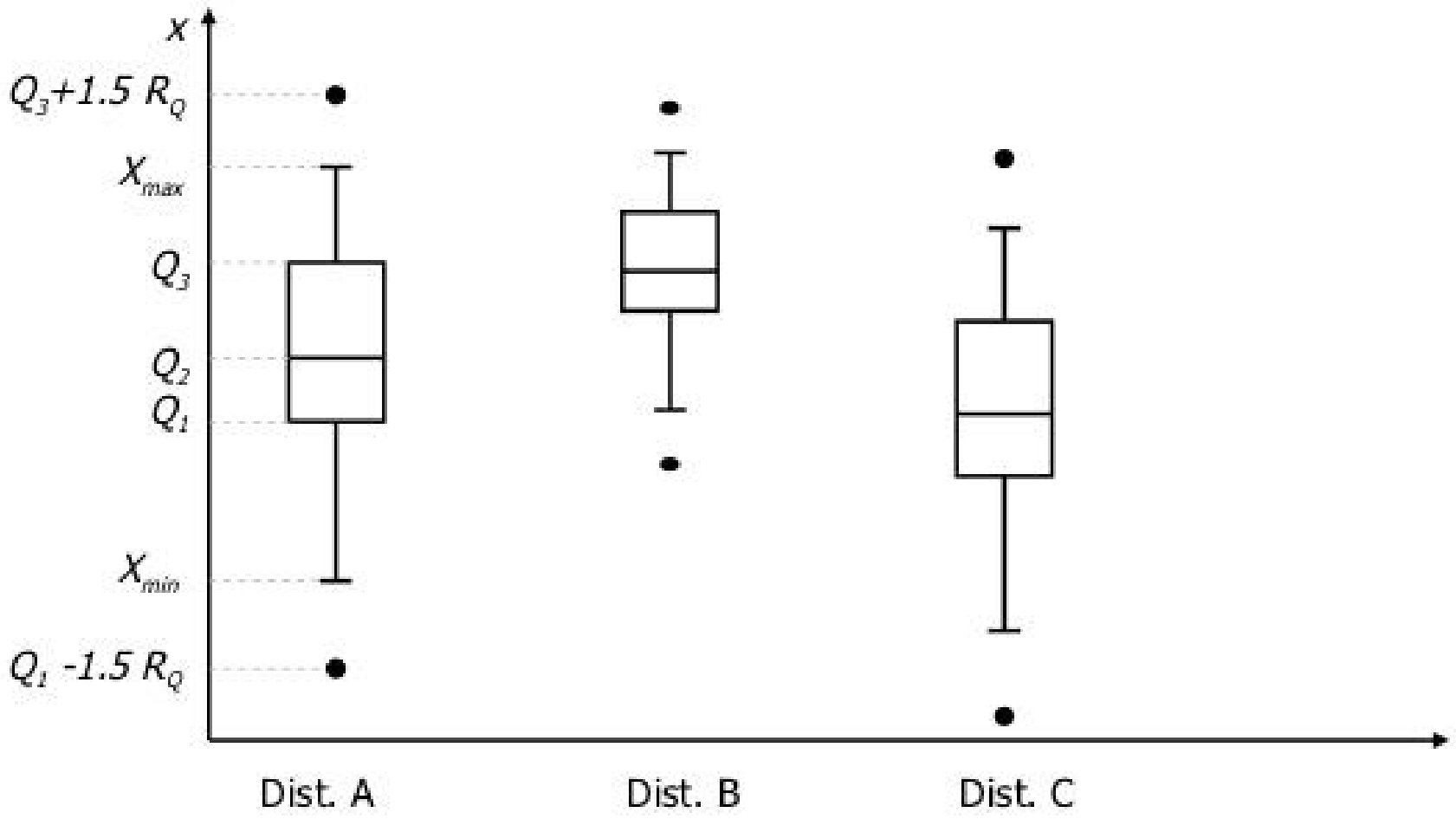
ESTADÍSTICA DESCRIPTIVA. INTRO



ESTADÍSTICA DESCRIPTIVA. INTRO



ESTADÍSTICA DESCRIPTIVA. INTRO



ESTADÍSTICA DESCRIPTIVA. INTRO

Estadísticos que podemos aplicar a los datos

- Datos Cuantitativos:
  - Varianza. Es un índice de dispersión de las variables.

    $$\sigma^2 = \Sigma(x_i - \bar{x})^2 / N$$

  - Desviación estándar. Es la raíz cuadrada de la varianza.
  - Coeficiente de dispersión: $\sigma / \bar{x}$

ESTADÍSTICA DESCRIPTIVA. INTRO

- Tanto la varianza como la desviación estándar tienen un problema. Lo vemos con un ejemplo:

  - Colectivo 1. Rinocerontes. Desviación estándar = 100 kilos
  - Colectivo 2. Perros. Desviación estándar = 100 kilos

- Viendo los datos anteriores, ¿podemos concluir que los datos son homogéneos?

$$\text{Peso rinocerontes} \approx 5000\text{Kg}$$
$$\text{Peso perros} \approx 30\ \text{Kg}$$

ESTADÍSTICA DESCRIPTIVA. INTRO

- Una desviación de 100 kg en los elefantes no es rara teniendo en cuenta que su peso medio es 5000Kg.
- Sin embargo, una desviación de 100Kg en perros es poco normal ya que de media pesan 30Kg, por tanto la distribución de los perros es más heterogénea que la de los elefantes. Necesitamos por tanto medir de alguna manera cómo de dispersa es la muestra.
- El estadístico a usar para medir esta dispersión se llama **coeficiente de dispersión** y se calcula como:

$$\sigma / \overline{x}$$

ESTADÍSTICA DESCRIPTIVA. INTRO

El problema es que la desviación estandar y la media DEBEN IR JUNTAS SIEMPRE, por si solas (ambas) NO DICEN NADA. Muestran una información parcial. Una variable necesita de la otra para ser interpretada.

TABLA DE CONTENIDOS

ESTADÍSTICOS BÁSICOS

- Media Aritmética.
  - La media x (también llamada promedio o media aritmética) de un conjunto de datos (X_1,X_2,...,X_N) es una medida de posición central. La definimos como el valor característico de la serie de datos resultado de la suma de todas las observaciones dividido por el número total de datos.

$$Media(X) = \overline{x} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

ESTADÍSTICOS BÁSICOS

- Media Aritmética.

➔

Estatura	Nº Personas n_i	M. Clase x_i	$n_i x_i$
140-150	20	145	2900
150-160	100	155	15500
160-180	80	170	13600
180-200	10	190	1900
	$n = 210$		33900

$$Media: \quad \bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} = \frac{33900}{210} = 161.42$$

ESTADÍSTICOS BÁSICOS

- Moda.
  - La moda es el valor de la variable más frecuente. Puede darse el caso de tener una muestra de datos, donde exista más de una moda (Plurimodal).

  - Moda para variables discretas

- Datos en serie
  2, 2, 3, 3, 3, 3, 5, 6, 7 $Mo = 3$

- Datos en tabla

♦ Ejemplo

x_i	n_i
1	34
2	36
3	45
4	22
5	17

$Mo = 3$

ESTADÍSTICOS BÁSICOS

- Moda.
  - Moda para variables continuas. La clase se corresponde con la mayor frecuencia (2,2)

$$Mo = e_{i-1} + \frac{h_i - h_{i-1}}{(h_i - h_{i-1}) + (hi - h_{i+1})} a_i$$

♦ **Ejemplo**

x_i	n_i	$h_i = n_i / a_i$
140-160	30	1.5
160-170	22	2.2
170-180	20	2
180-190	18	1.8
190-200	10	1
	100	

$$Mo = 160 + \frac{(2.2 - 1.5)}{(2.2 - 1.5) + (2.2 - 2)} \times 10 = 167.777$$

e_{i-1}= Límite inferior de la clase modal

a_i= Amplitud del intervalo

h_i-h_{i-i}= Diferencia entre la frecuencia absoluta de la clase modal y la de la clase anterior

h_i-h_{i+i}= Diferencia entre la frecuencia absoluta de la clase modal y la de la clase siguiente

ESTADÍSTICOS BÁSICOS

- Mediana.
  - Valor de la variables que ocupa el lugar central de usa serie de datos ordenados. Divide la distribución de los datos en 2 partes iguales.
  - El 50% de los elementos de la población tienen un valor de la variable menor que la mediana y por tanto el 50% de los elementos restantes, tendrán un valor de la variable mayor que la mediana.

ESTADÍSTICOS BÁSICOS

- Mediana.

➢ **Variables discretas**

▪ Datos en tabla

♦ **Ejemplo**

x_i	n_i	N_i	f_i	F_i
0	4	4	0.142	0.142
1	6	10	0.214	0.357
2	10	20	0.357	0.714
3	5	25	0.178	0.892
4	3	28	0.107	1
	28		1	

$n/2 = 14$
$F_i = 1/2$

$Me = 2$

➢ **Variables continuas**

$$Me = e_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i = e_{i-1} + \frac{\frac{1}{2} - F_{i-1}}{f_i} a_i$$

♦ **Ejemplo**

Tallas	n_i	N_i	f_i	F_i
140-150	15	15	0.15	0.15
150-160	30	45	0.30	0.45
160-170	25	70	0.25	0.70
170-180	20	90	0.20	0.90
180-200	10	100	0.10	1
	100			

$n/2 = 50$
$F_i = 1/2$

$$Me = 160 + \frac{0.5 - 0.45}{0.25} \times 10 = 160 + 2 = 162$$

ESTADÍSTICOS BÁSICOS

- Percentiles
  - P_k, k=1,2,...,99. Percentil k, es el valor de la variable que deja por debajo, el k% de los valores de la variable.

$$Q_1 = P_{25} \rightarrow \text{Cuartil } 1^o$$
$$Q_2 = P_{50} \rightarrow \text{Cuartil } 2^o = Me$$
$$Q_3 = P_{75} \rightarrow \text{Cuartil } 3^o$$

⇐ Mediana

$$D_1 = P_{10} \rightarrow \text{Decil } 1^o$$
$$D_2 = P_{20} \rightarrow \text{Decil } 2^o$$
$$\ldots$$
$$D_9 = P_{90} \rightarrow \text{Decil } 9^o$$

ESTADÍSTICOS BÁSICOS

- Percentiles

- Cálculo para v.e. discretas:

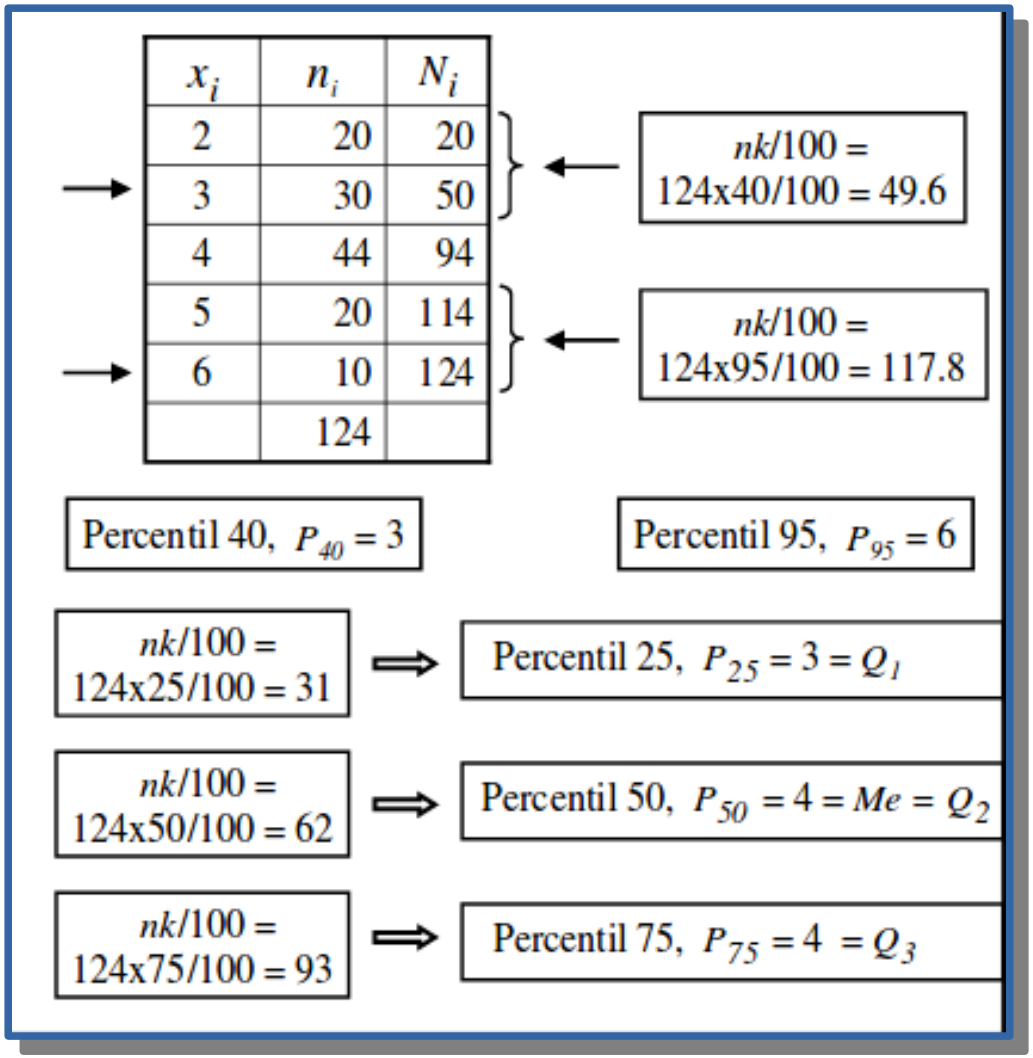
Igual que la mediana, cambiando $n/2$ por $nk/100$

- Cálculo para v.e. continuas:

$$P_k = e_{i-1} + \frac{\frac{nk}{100} - N_{i-1}}{n_i} a_i = e_{i-1} + \frac{\frac{k}{100} - F_{i-1}}{f_i} a_i$$

ESTADÍSTICOS BÁSICOS

- Percentiles. Ejemplo con variables estadísticas discretas



ESTADÍSTICOS BÁSICOS

- Percentiles. Ejemplo con variables estadísticas contínuas

Tallas	n_i	N_i	f_i	F_i	
140-150	15	15	0.15	0.15	
150-160	30	45	0.30	0.45	P_{40}
160-170	25	70	0.25	0.70	
170-180	20	90	0.20	0.90	P_{75}
180-200	10	100	0.10	1	
	100				

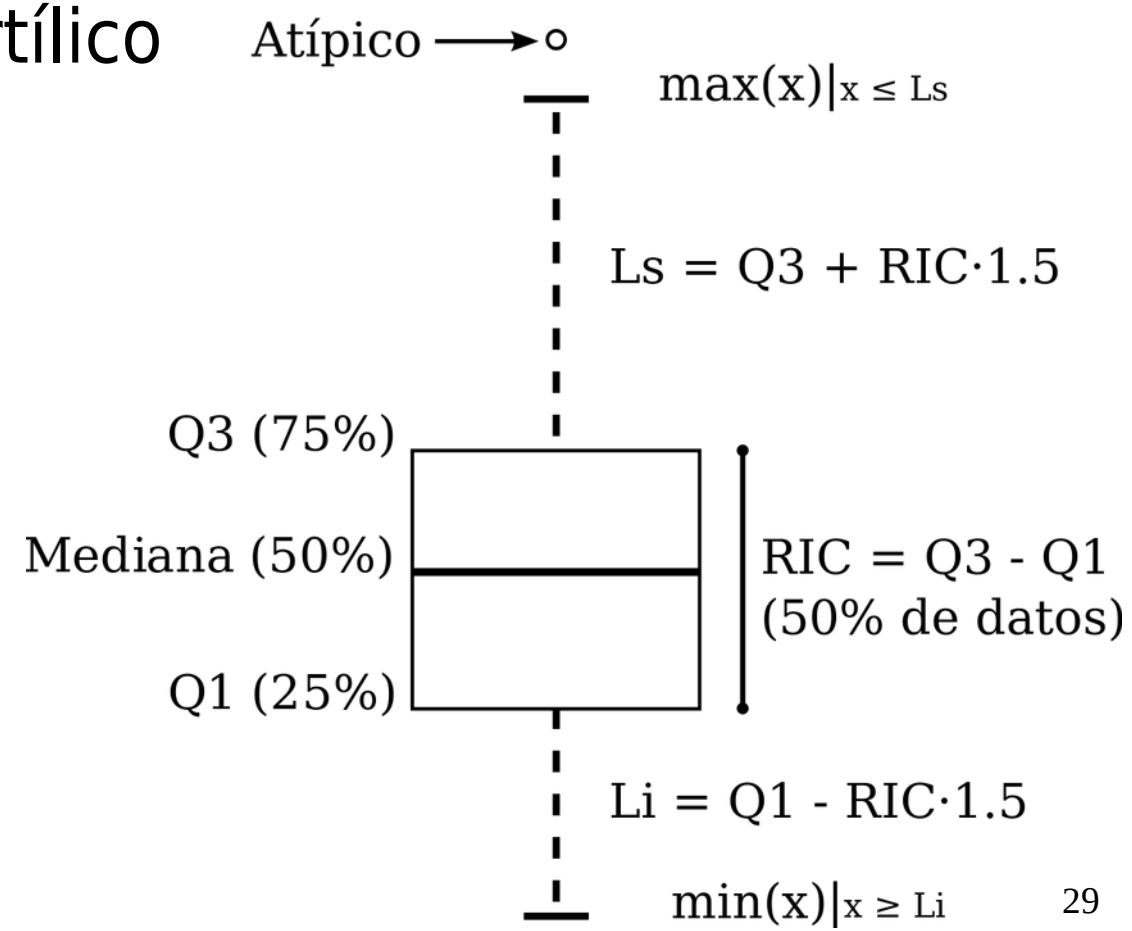
$$P_k = e_{i-1} + \frac{\frac{nk}{100} - N_{i-1}}{n_i} a_i = e_{i-1} + \frac{\frac{k}{100} - F_{i-1}}{f_i} a_i$$

$$P_{40} = 150 + \frac{40-15}{30} \times 10 = 150 + \frac{0.4-0.15}{0.30} \times 10 = 158.33$$

$$P_{75} = 170 + \frac{75-70}{20} \times 10 = 170 + \frac{0.75-0.70}{0.20} \times 10 = 172.5 = Q_3$$

ESTADÍSTICOS BÁSICOS

- Rango o recorrido. Valor máximo menos el valor mínimo de la variable.
- Recorrido intercuartílico
  - Q_3-Q_1



ESTADÍSTICOS BÁSICOS

- Varianza. Es una medida vinculada a la dispersión de una variable. A mayor dispersión, mayor variabilidad.

$$\sigma^2 = \frac{\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^2}{n} = \frac{\sum_{i=1}^{k} n_i x_i^2}{n} - \bar{x}^2$$

- Desviación típica. Se define como la raíz cuadrada de la varianza.

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

ESTADÍSTICOS BÁSICOS

- Coeficiente de variación.

$$C.V. = \frac{\sigma}{\bar{x}}$$

- Coeficiente de correlación (entre variables). Es una medida de regresión que cuantifica el grado de variación conjunta entre 2 variables cuantitativas.

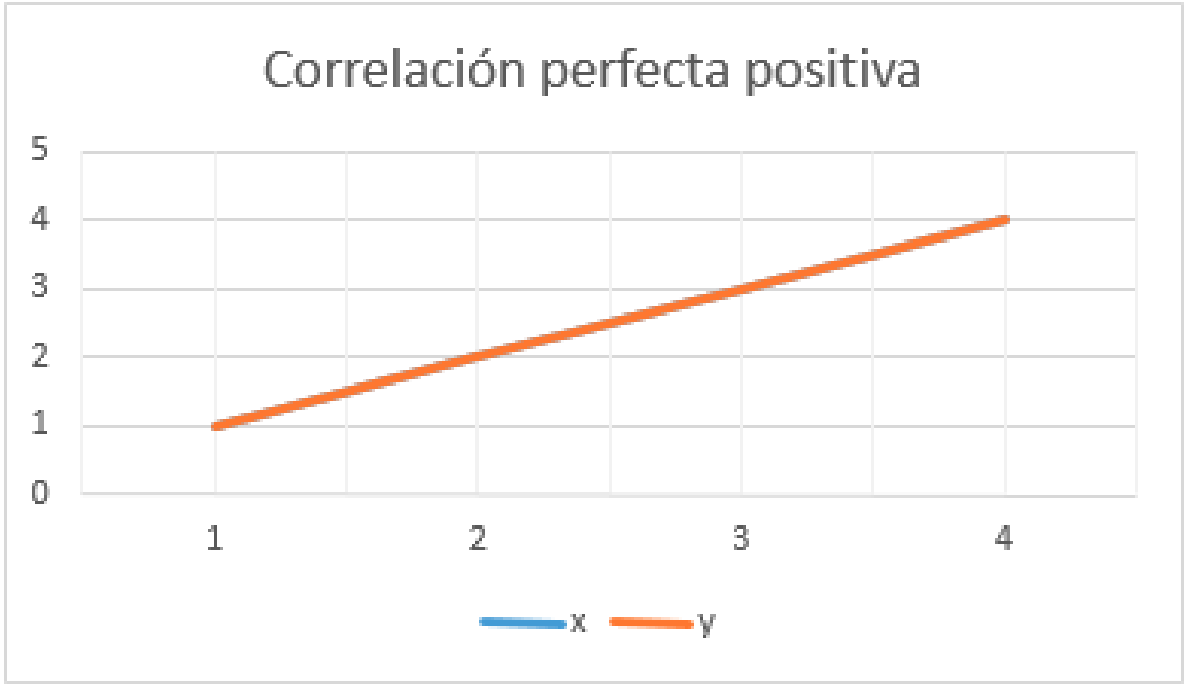
ESTADÍSTICOS BÁSICOS

- La correlación puede tomar valores entre -1 y 1
  - R^2 = -1, correlación perfecta negativa entre variables
  - R^2 = 0, no existe correlación entre variables
  - R^2 = 1, correlación perfecta positiva entre variables

ESTADÍSTICOS BÁSICOS

- Correlación perfecta positiva:

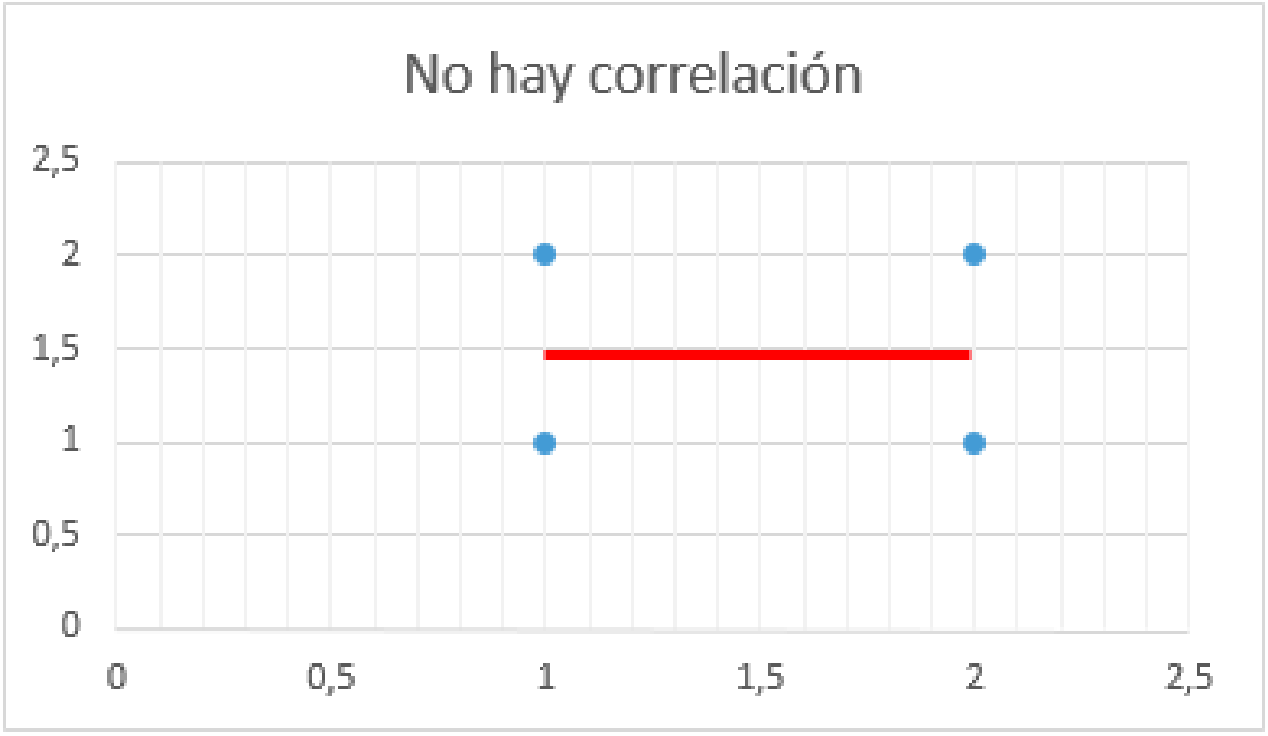
x	y
1	1
2	2
3	3
4	4
Correlación	**1**



ESTADÍSTICOS BÁSICOS

- Sin correlación:

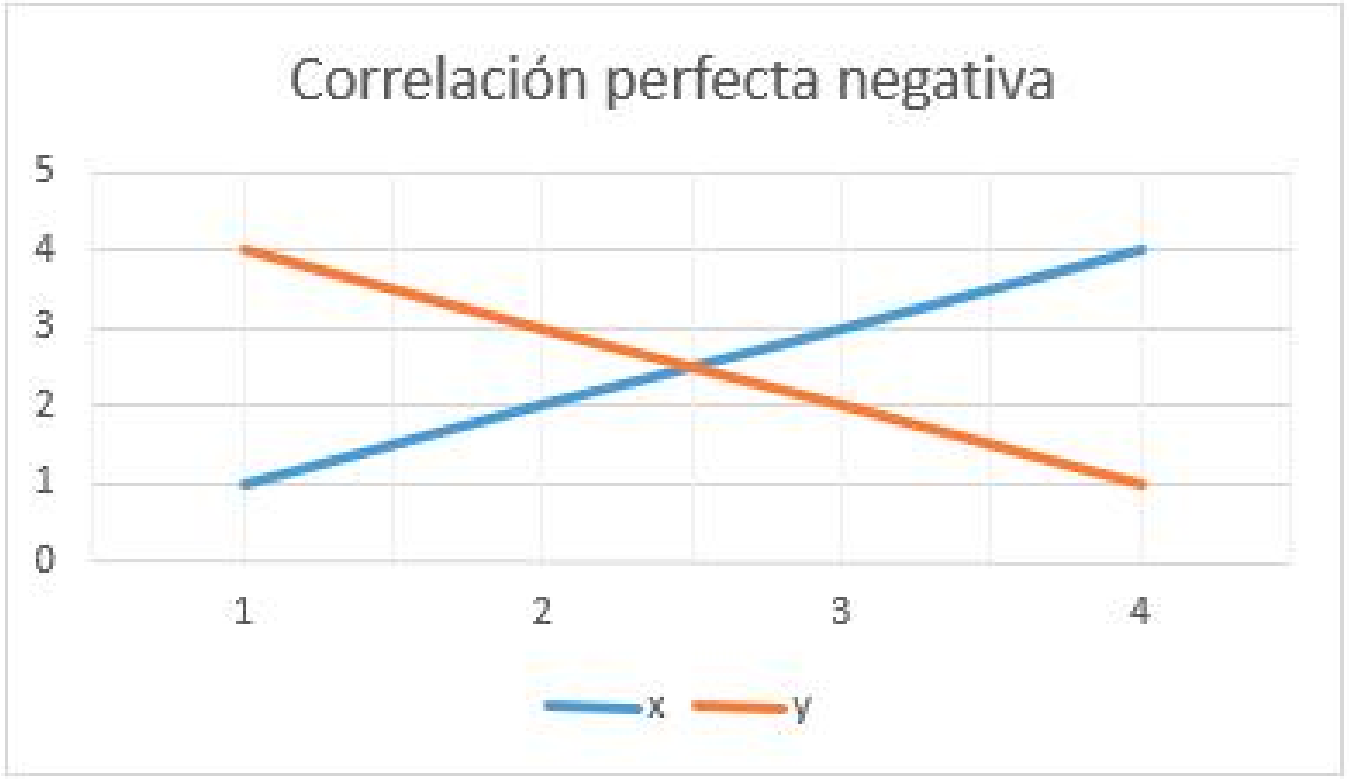
x	y
1	1
2	1
2	2
1	2
Correlación	**0**



ESTADÍSTICOS BÁSICOS

- Correlación perfecta negativa:

x	y
1	4
2	3
3	2
4	1
Correlación	-1



ESTADÍSTICOS BÁSICOS

- Cálculo R^2

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

- Donde cov_{xy} (covarianza entre x e y)

$$Cov[X,Y] = \sigma_{xy} = \frac{\sum_i \sum_j n_{ij} (x_i - \bar{x})(y_j - \bar{y})}{n} =$$

$$= \frac{\sum_i \sum_j n_{ij} x_i y_j}{n} - \bar{x}\,\bar{y}$$

TABLA DE CONTENIDOS

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Vamos a estudiar como se relacionan las variables entre si.
- Para los modelos de predicción que vamos a utilizar durante el curso, necesitamos conocer la relación entre las variables.
- Las variables que usemos, deben estar relacionadas entre si.
    - *No tiene sentido por ejemplo, querer predecir el peso de una persona a partir del color de sus ojos, pero sí a partir del número de calorías consumidas, o del nivel de actividad.*

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Para simplificar, para 2 variables y en función de su naturaleza, podemos encontrarnos con 3 tipos de relaciones:
  1. Que las 2 sean cualitativas.
  2. Que una sea cualitativa y otra cuantitativa.
  3. Que las 2 sean cuantitativas.

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- **Ambas variables son cualitativas.**
  - Supongamos que tenemos 2 variables, “clase social” y “barrio de residencia”.
  - Cada variable contiene los siguientes estados:
    - Clase social → Baja, Media, Alta
    - Barrio Residencia → A, B, C

Clase Social	Baja
	Media
	Alta
Barrio	A
	B
	C

	Clase Social	Barrio
1	B	C
2	A	A
3	M	B
…	…	…
1000	A	B

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Para poder conocer la relación entre estas variables, necesitamos crear una **tabla de contingencia** que incluya todas las posibles combinaciones.

	Baja	Media	Alta
A			
B			
C			

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Si están asociadas/relacionadas, conocer una me coloca en mejor posición para predecir la otra.
- Completamos la tabla de contingencia con la cantidad de veces que se da cada par de datos:

  A y Baja, A y Media, A y Alta

  ...

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Básicamente estamos rellenando la tabla con las **FRECUENCIAS ABSOLUTAS DE CADA CRUCE.**
- De esta manera obtenemos la **distribución de frecuencias conjuntas** de las variables

Modelo observado

Valores absolutos

1	BAJA	MEDIA	ALTA	
A	50	100	50	200
B	50	50	200	300
C	400	50	50	500
	500	200	300	1000

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

Modelo observado

Valores absolutos

1	BAJA	MEDIA	ALTA	
A	50	100	50	200
B	50	50	200	300
C	400	50	50	500
	500	200	300	1000

- La zona azul, muestra la distribución de frecuencias POR BARRIO
- La zona verde, muestra la distribución de frecuencias POR CLASE SOCIAL

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- A simple vista, podemos ver que existe relación entre la clase social y el barrio de residencia, pero podría no verse tan claro.
- Para poder verlo mejor, podemos mostrar las frecuencias relativas respecto del total del barrio o de la clase.
  - *Ej, frecuencias relativas respecto del barrio*

Para A,Baja = 0,25 → 50/200= 0,25
Para B, Baja = 0,17 → 50/300 = 0,17
Para C, Baja = 0,80 → 400/500 = 0,80

Modelo observado

Valores relativos por barrio

2

	BAJA	MEDIA	ALTA	
A	0,25	0,50	0,25	1,00
B	0,17	0,17	0,67	1,00
C	0,80	0,10	0,10	1,00

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Supongamos que la proporción de clases es la misma que la de la tabla 2.
- Pregunta para nota. Si la clase social y el barrio de residencia fueran independientes entre si (no tuvieran nada que ver), ¿qué números tendríamos en la tabla?

Modelo observado

Valores relativos por barrio

2

	BAJA	**MEDIA**	**ALTA**	
A	0,25	0,50	0,25	1,00
B	0,17	0,17	0,67	1,00
C	0,80	0,10	0,10	1,00
	0,50	0,20	0,30	

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Si la distribución de las clases sociales de un colectivo en un país fuera:
  - Baja 50%, Media 20%, Alta 30%
- Y las variables clase social y barrio de residencia fueran independientes, entonces en cada barrio, DEBERÍA HABER LA MISMA DISTRIBUCIÓN DE LOS DATOS

Modelo esperado

Lo que esperamos si los barrios fueran independientes.

3

	BAJA	MEDIA	ALTA	
A	100	40	60	200
B	150	60	90	300
C	250	100	150	500
	500	200	300	1000

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

*Para A,Baja = 100 → 200*500/1000*

*Para B, Baja = 150 → 300*500/1000*

*Para C, Baja = 250→ 500*500/1000*

*Para A, Media = 40 → 200*200/1000*

*Para B, Media = 60 → 300*200/1000*

*Para A, Alta = 60 → 300*300/1000*

*Para C, Alta = 150 → 500*300/1000*

Modelo observado

Valores absolutos

1

	BAJA	MEDIA	ALTA	
A	50	100	50	200
B	50	50	200	300
C	400	50	50	500
	500	200	300	1000

Modelo observado

Valores relativos por barrio

2

	BAJA	MEDIA	ALTA	
A	0,25	0,50	0,25	1,00
B	0,17	0,17	0,67	1,00
C	0,80	0,10	0,10	1,00
	0,50	0,20	0,30	

Modelo esperado

Lo que esperamos si los barrios fueran independientes.

3

	BAJA	MEDIA	ALTA	
A	100	40	60	200
B	150	60	90	300
C	250	100	150	500
	500	200	300	1000

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Hasta ahora tenemos 3 tablas, aunque para llegar a la tabla 3 podemos saltarnos la tabla 2.

Modelo observado
Valores absolutos

1

	BAJA	MEDIA	ALTA	
A	50	100	50	200
B	50	50	200	300
C	400	50	50	500
	500	200	300	1000

Modelo observado
Valores relativos por barrio

2

	BAJA	MEDIA	ALTA
A	0,25	0,50	0,25
B	0,17	0,17	0,67
C	0,80	0,10	0,10
	0,50	0,20	0,30

Modelo esperado
Lo que esperamos si los barrios fueran independientes.

3

	BAJA	MEDIA	ALTA	
A	100	40	60	200
B	150	60	90	300
C	250	100	150	500
	500	200	300	1000

- La tabla 3 es la que esperaríamos encontrarnos SÍ y SOLO SÍ, las variables Barrio y Clase Social no tuvieran nada que ver una con la otra.

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Computacionalmente pasar de la tabla 1 a la tabla 3 es muy fácil.
- Para saber si 2 variables son independientes entre sí o no, tendríamos que comprar los modelos 1 y 3.
- **¿En qué grado son dependientes o independientes?**

- Haremos uso de una nueva tabla llamada **MODELO DE RESIDUOS BRUTOS** que consiste en restar a lo observado, lo esperado.

Modelo de residuos brutos

Observado - esperado

4

	BAJA	MEDIA	ALTA	
A	-50	60	-10	0
B	-100	-10	110	0
C	150	-50	-100	0
	0	0	0	50

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Cuanto más se acerquen los valores a 0, **MENOS** relacionadas estarán las variables →

  El residuo es menor y por tanto más se acerca lo observado a lo esperado.

- Esta última tabla no nos muestra información muy clara de lo dependientes que son las variables entre si.
- Usaremos una quinta tabla:

Residuo Estandarizado = Residuo Bruto / √ (Modelo Esperado)

Residuo estandarizado

Residuo bruto / Raiz(esperado)

5

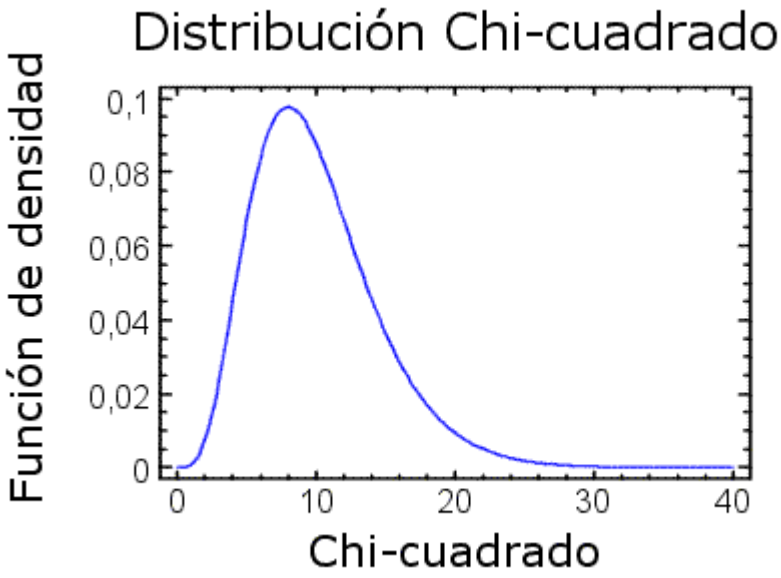
	BAJA	MEDIA	ALTA
A	-5,00	9,49	-1,29
B	-8,16	-1,29	11,60
C	9,49	-5,00	-8,16

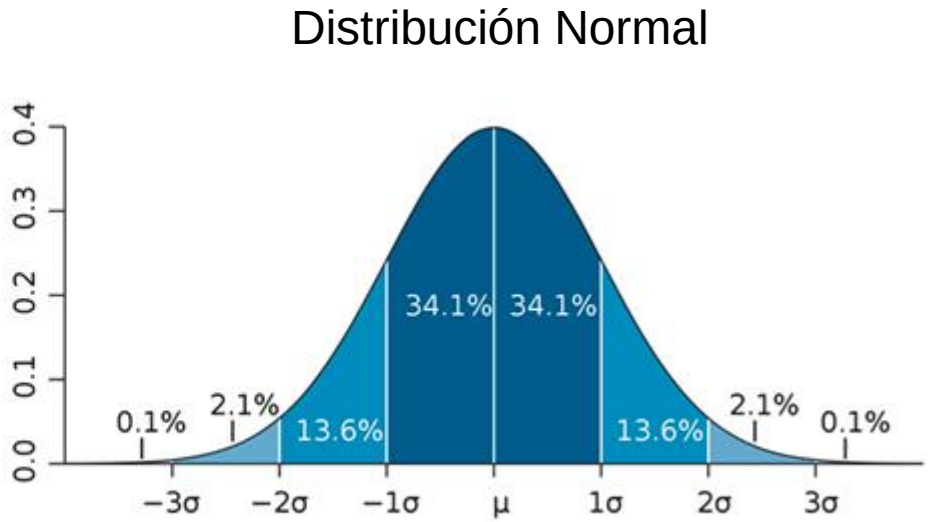
Para A,Baja = -5 → -50 / √100
Para B, Baja = 8,16 → -100 / √150
Para C, Baja = 9,49→ 150 / √250

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Con la quinta tabla, lo que estamos haciendo es RELATIVIZAR los datos respecto de algo.

Al dividir los residuos brutos entre la raíz del residuo bruto esperado, hacemos que la distribución se ajuste a una distribución Chi Cuadrado.





TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Ahora sumamos todos los residuos estandarizados al cuadrado y cuanto más se aleje el resultado de 0, MÁS ASOCIACIÓN EXISTIRÁ ENTRE LAS VARIABLES.

6	**Suma cuadrados residuos estandarizados** 501,11

- ¿Cuánto es, cuanto más se aleje?¿A partir de qué punto podemos decir que existe relación entre las variables?

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- **Estadísticamente**, el punto a partir de cual consideramos que unas variables están asociadas (el valor crítico) **depende del tamaño de la tabla y un nivel de exigencia**.
- El tamaño de la tabla define los grados de libertad (dato objetivo) → conocemos este dato a priori.
- Al sumar elementos al cuadrado, el tamaño de la tabla influye en el valor obtenido.
- **El nivel de exigencia, es un valor subjetivo y se conoce también como nivel de significación (α).**
- El nivel de exigencia depende de la persona, es subjetivo.

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Haciendo un simil jurídico y el nivel de significación →

  A una persona se le procesa por defraudar a hacienda. Podría darse el caso de que un juez exija muchas evidencias para condenar a una persona, mientras que otro juez no.

- Ese nivel de evidencias es lo que estadísticamente se entiende nivel de significación o exigencia.
- Para 2 personas, el nivel de exigencia para considerar algo como cierto puede variar.
- Una persona puede considerar que con un nivel de residuos menor que otra, las variables ya pueden considerarse como independientes, mientras que otra puede exigir más evidencias.

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Los grados de libertad en una tabla, se definen como:

(Filas – 1) *	**(Columnas – 1)**	**´= Grados de libertad**
(3-1) *	(3-1)	´= 4

- Los valores habituales de nivel de significación (exigencias a la evidencia) son:
  - 0,01 → 99% de probabilidad de que los resultados del análisis sean ciertos.
  - 0,05 → 95% de probabilidad de que los resultados del análisis sean ciertos.
  - 0,10 → 90% de probabilidad de que los resultados del análisis sean ciertos.

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- La distribución Chi Cuadrado (χ^2)

Table 3-1 Critical Values of the χ^2 Distribution

	P								
df	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Table 3-1
Introduction to Genetic Analysis, Tenth Edition

TÉCNICAS DE MEDICIÓN/ASOCIACIÓN ENTRE VARIABLES

- Siguiendo nuestro ejemplo, para un $\alpha = 0,01$, el resultado sería 13,277. Para un $\alpha = 0,05$, el resultado sería 9,488 y para un $\alpha = 0,1$ sería 7,779
- Esto quiere decir que para cualquiera de los α que hemos tomado, el valor crítico de Chi cuadrado es menor ($13,277 < 501,11$, $9,488 < 501,11$, $7,779 < 501,11$) que la suma de los cuadrados de los residuos estandarizados.
- Por tanto podríamos concluir **para nuestro caso**, que existe relación entre las variables para cualquiera de los 3 niveles de significación tomados.

TABLA DE CONTENIDOS

INFERENCIA ESTADÍSTICA

- 2 tipos de aprendizaje estadístico:
  1) Aprendizaje supervisado. Para predecir / estimar cosas.
  2) Aprendizaje no supervisado. Encontrar estructuras / patrones donde a priori no las hay.

- Aprendizaje no supervisado → Clustering campañas márketing.
- Aprendizaje supervisado → Medición del % de grasa corporal. Variables Proxy

INFERENCIA ESTADÍSTICA

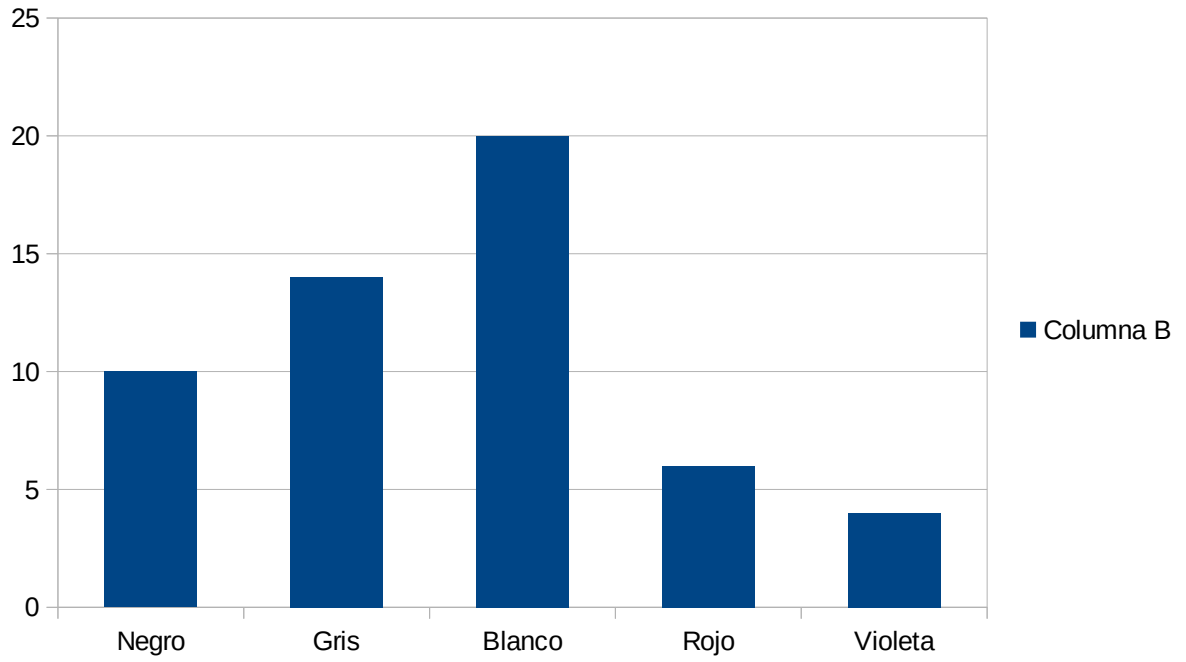
- El problema de los modelos supervisados está en el OVERFITTING = sobreentrenamiento de los modelos.
- Si sobreentrenamos el modelo, éste se ajustará demasiado bien a la muestra de aprendizaje e incorporará en el modelo las particularidades propias de ese conjunto de datos, más allá de lo que son las verdaderas relaciones entre las variables explicativas y la variable a predecir.
- Si esto ocurre, estaré ajustando ruido en vez de señal.
- Esto pasa cuando creamos modelos con funciones muy complejas, o con muchas variables, ...

TABLA DE CONTENIDOS

GRÁFICOS ESTADÍSTICOS

- Variables estadísticas Cualitativas. Gráficos de Barras

Color Plumaje	N.º de Aves (n_i)
Negro	10
Gris	14
Blanco	20
Rojo	6
Violeta	4



GRÁFICOS ESTADÍSTICOS

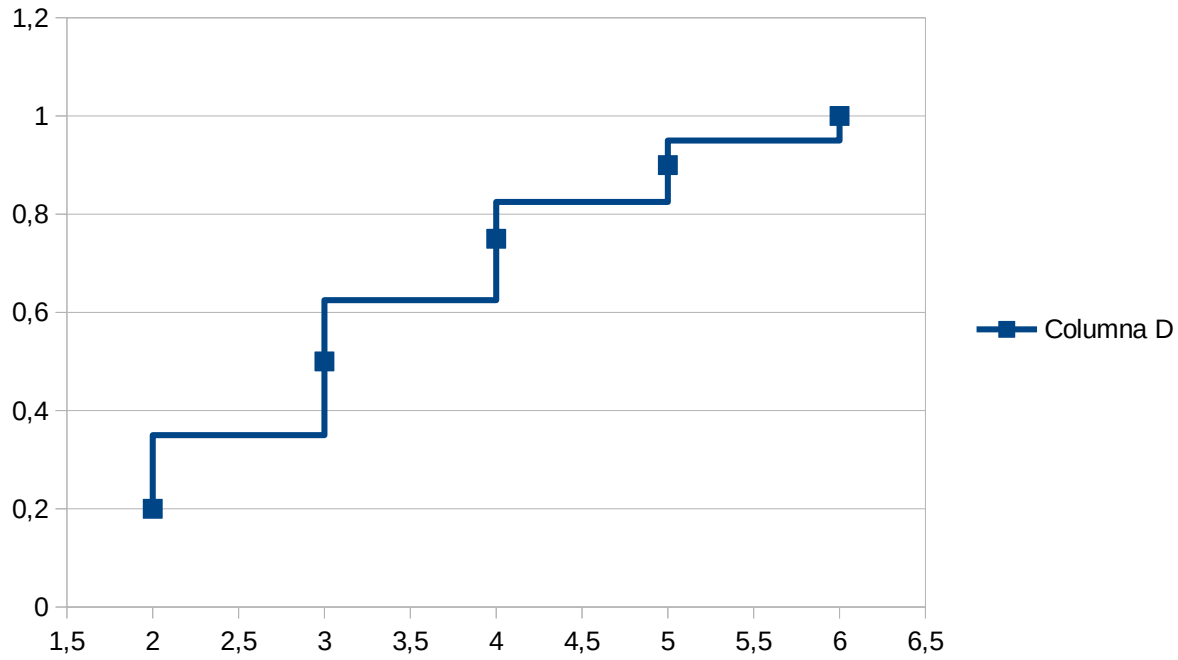
- Variables estadísticas Cualitativas. Gráficos de Sectores

Color Plumaje	N.º de Aves (n_i)
Negro	10
Gris	14
Blanco	20
Rojo	6
Violeta	4

GRÁFICOS ESTADÍSTICOS

- Variables estadísticas Discretas. Curva Acumulativa de distribución

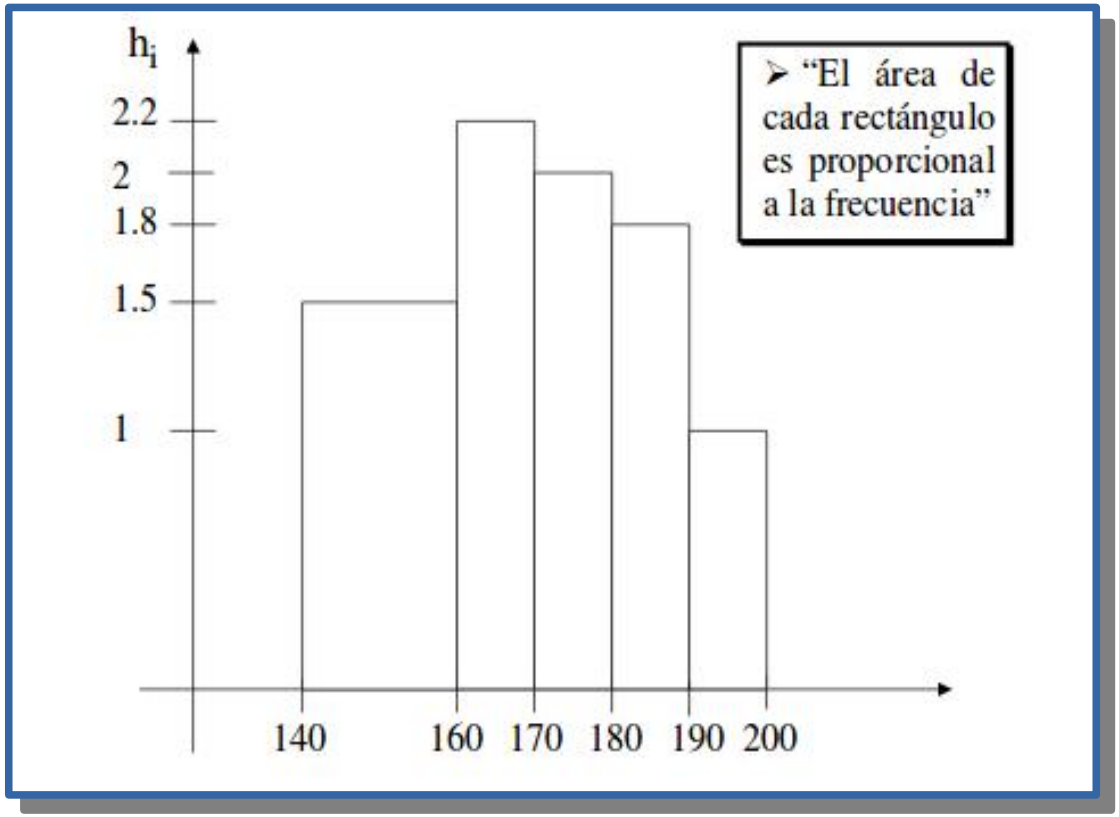
N.º de crías	N.º de animales : n_i	f_i	F_i
2	20	0,20	0,20
3	30	0,30	0,50
4	25	0,25	0,75
5	15	0,15	0,90
6	10	0,10	1
	n=100		



GRÁFICOS ESTADÍSTICOS

- Variables estadísticas Continuas Histograma.

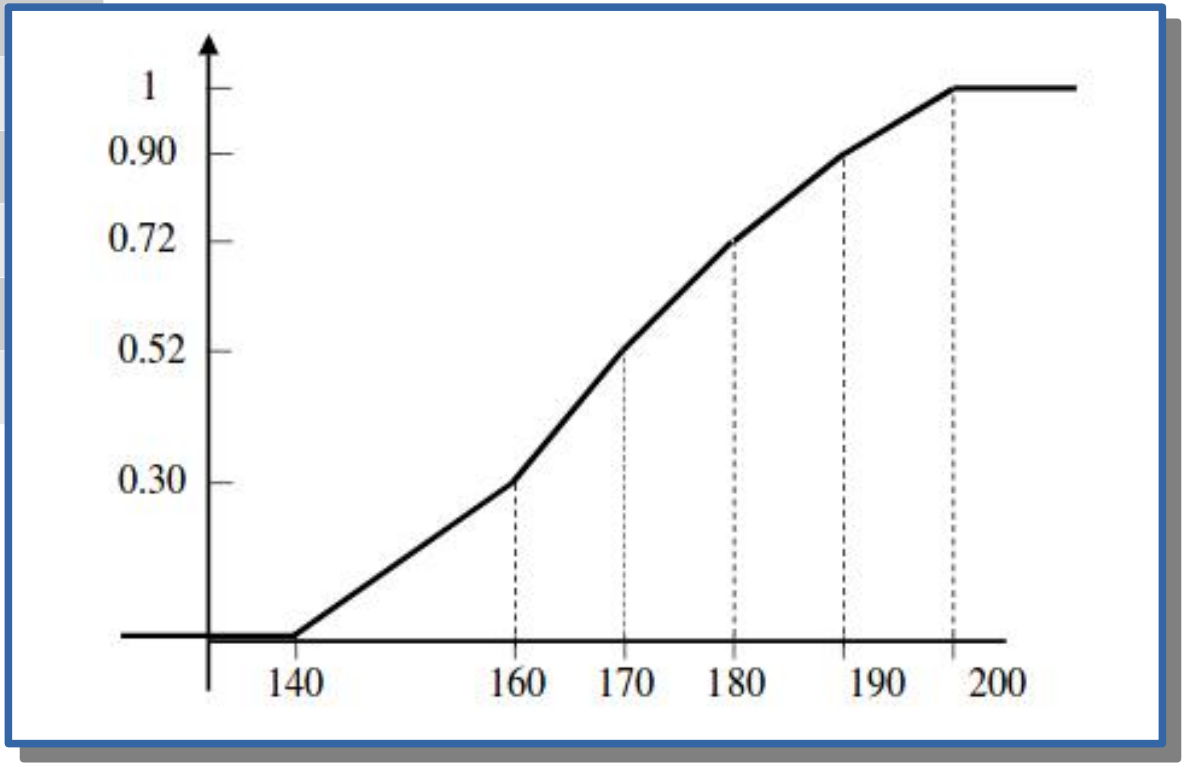
Estatura	**n_i**	**$h_i=n_i / a_i$**
140-160	30	1,5
160-170	22	2,2
170-180	20	2
180-190	18	1,8
190-200	10	1
	100	



GRÁFICOS ESTADÍSTICOS

- Variables estadísticas. Curva acumulativa de distribución.

Estatura	n_i	f_i	F_i
140-160	30	0,3	0,3
160-170	22	0,22	0,52
170-180	20	0,2	0,72
180-190	18	0,18	0,9
190-200	10	0,1	1
	100		



RECURSOS

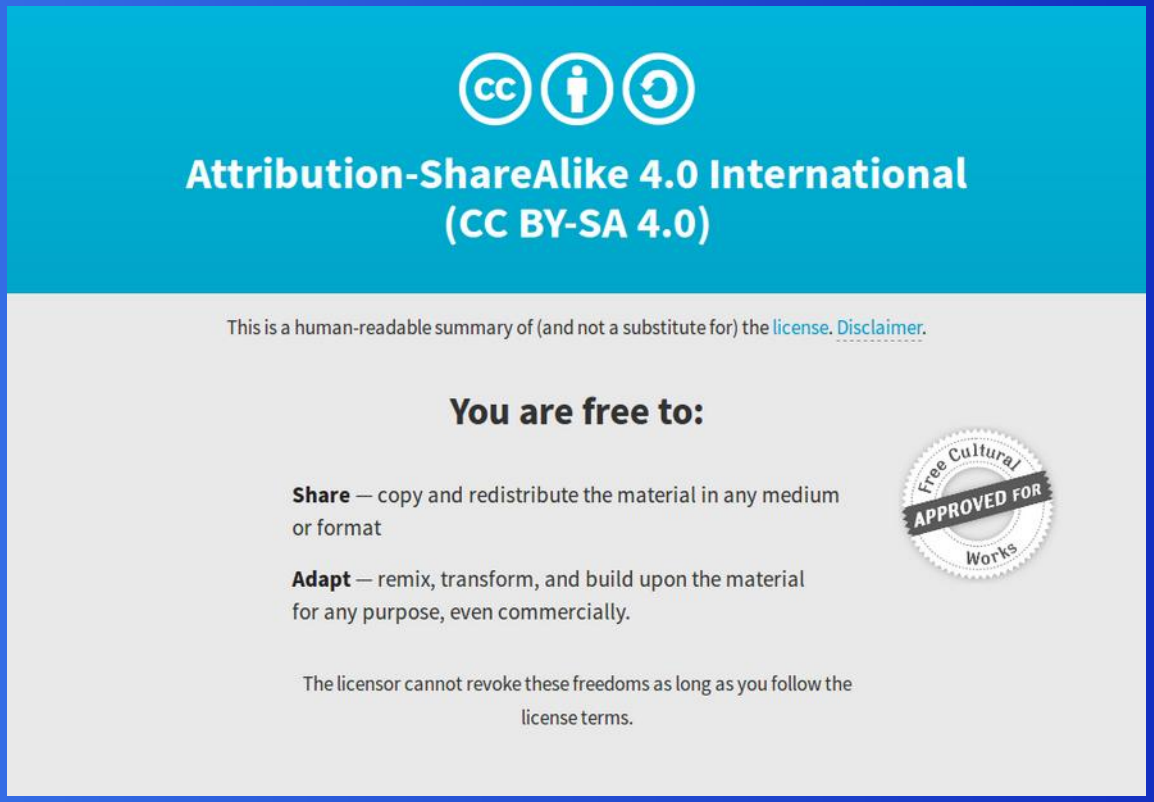
- Recursos sobre estadística descriptiva:

  http://132.248.164.227/publicaciones/docs/apuntes_matematicas/34.%20Estadistica%20Descriptiva.pdf

  http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf

- Recursos sobre gráficos en Python:

  https://www.python-graph-gallery.com/



Germán Alonso Lascurain
germanalonso@opendeusto.es

CÁMARABILBAO C2B campus to business ai aipower tigloo