BLOQUE I
ENTENDIENDO Y EXPLORANDO LOS DATOS

PEDRO M. VALERO MORA

GRUPO B, G Y H

CURSO 2009-2010

Parte I
Datos

1.1.Ejemplo de datos

"In God we Trust...All Others Bring Data"
(W.E. Demming)

Ejemplo

En la página del curso podeis encontrar un link para un archivo de datos llamado Bycicles.sav. En ese link teneis un archivo de datos sobre una investigación realizada en la universidad de Bath (para más información http://www.drianwalker.com/work.html).En estos datos tenemos los resultados de un estudio en el que el autor estudió si el llevar casco de bicicleta, si el llevar peluca de mujer, si el tipo de vehículo que se trate, afectan a la distancia con que otros vehículos adelantan a los que

van en bic

1.2.Qué son datos

- Para que unos números sean algo más que datos necesitamos responder a las 6 Q.
  - Quién
  - Qué
  - Qomo, Quando, por Que y Donde
- Veamos esto con más detalle

1.3.Quién

Observaciones

- Echemosle un vistazo a los datos de las bicicletas

	vehicle	Colour	hour	helmet	Distance_from_kerb	Passing_distance	root	va
1	car	blue	16:00:00	0	.5	1.864	1.37	
2	HGV	red	16:00:00	0	.5	.748	.86	
3	LGV	blue	16:00:00	0	.5	1.567	1.25	
4	car	unknown	16:00:00	0	.5	1.390	1.18	
5	bus	other	16:00:00	0	.5	1.294	1.14	
6	car	silver/grey	16:00:00	0	.5	1.259	1.12	
7	LGV	white	7:00:00	1	1.0	1.040	1.02	
8	car	red	7:00:00	1	1.0	1.262	1.12	
9	bus	red	7:00:00	1	1.0	.799	.89	
10	PTW	unknown	7:00:00	1	1.0	1.682	1.30	

- En general quien hace referencia a los individuos que hay en los datos. Esto puede hacer referencia a:
  - Encuestados (sujetos de una encuesta)
  - Sujetos o participantes (en un experimento)
  - Unidades experimentales (en un experimento que no involucre sujetos humanos)
  - Registros (en una base de datos)
  - Observaciones (más general y se puede acoplar al caso que estamos viendo)
  - Casos (también bastante general)

- Lo más importante a recordar es que normalmente ***los casos se ponen en las filas de la tabla de datos***
  - Es decir, mirando a lo largo de una fila tenemos los valores que un sujeto/participante/observación/caso/ unidad experimental/registro/caso tiene

1.4.Qué

- Las características que se registran de cada individuo se llaman variables

	vehicle	Colour	hour	helmet	Distance_from_kerb	Passing_distance	root	va
1	car	blue	16:00:00	0	.5	1.864	1.37	
2	HGV	red	16:00:00	0	.5	.748	.86	
3	LGV	blue	16:00:00	0	.5	1.567	1.25	
4	car	unknown	16:00:00	0	.5	1.390	1.18	
5	bus	other	16:00:00	0	.5	1.294	1.14	
6	car	silver/grey	16:00:00	0	.5	1.259	1.12	
7	LGV	white	7:00:00	1	1.0	1.040	1.02	
8	car	red	7:00:00	1	1.0	1.262	1.12	
9	bus	red	7:00:00	1	1.0	.799	.89	
10	PTW	unknown	7:00:00	1	1.0	1.682	1.30	

  - Las variables se ponen en las columnas

- Hay dos tipos de variables fundamentales
  - Categóricas (aunque se use números, los números no son propiamente números sino que representan categorías)
  - Numéricas (los números son números de verdad)

- Hay un tercer tipo que está en medio
  - Las variables ordinales están a mitad de camino entre las variables categóricas y las numéricas
  - A veces se parecen más a una variable categórica (por ejemplo, ¿Cree que las patatas fritas deben estar aceitosas? 1:A favor; 2:Indiferente; 3: En contra)
  - Si le ponemos más valores a una variable ordinal entonces se parece más a una variable numérica: (Valora de 1 a 100 la cantidad de aceite que debe haber en una patata frita para que la experiencia gastronómica derivada de su degustación alcance las cimas más elevadas)

- Las variables ordinales son problemáticas así que es conveniente estar atento cuando aparezcan en la asignatura ya que es difícil dar reglas generales sobre su uso y a menudo se usan técnicas específicas
- En cualquier caso, el lugar más adecuado para discutir largo y tendido sobre el tema es en la asignatura de Psicometría, no en Análisis de Datos

- Un último tipo de variable son las binarias o dicotómicas
  - En el ejemplo de las bicicletas, la variable casco (helmet) es de ese tipo
  - Este tipo de variables se puede usar correctamente tanto como numérica o como categórica

1.5.Dónde, Cuándo, Cómo y Por qué

El contexto de los datos

- Los datos necesitan contexto para que tengan significado.
  - Dónde y Cuándo: Datos registrados en un sitio pueden tener un significado diferente a datos registrados en un sitio diferente. Lo mismo con el tiempo o época en que fueron registrados
  - Cómo: Vosotros teneis una asignatura de métodos de investigación que os enseña cómo recoger unos datos adecuadamente. Según como estén recogidos los datos, las interpretaciones o consecuencias están más limitadas

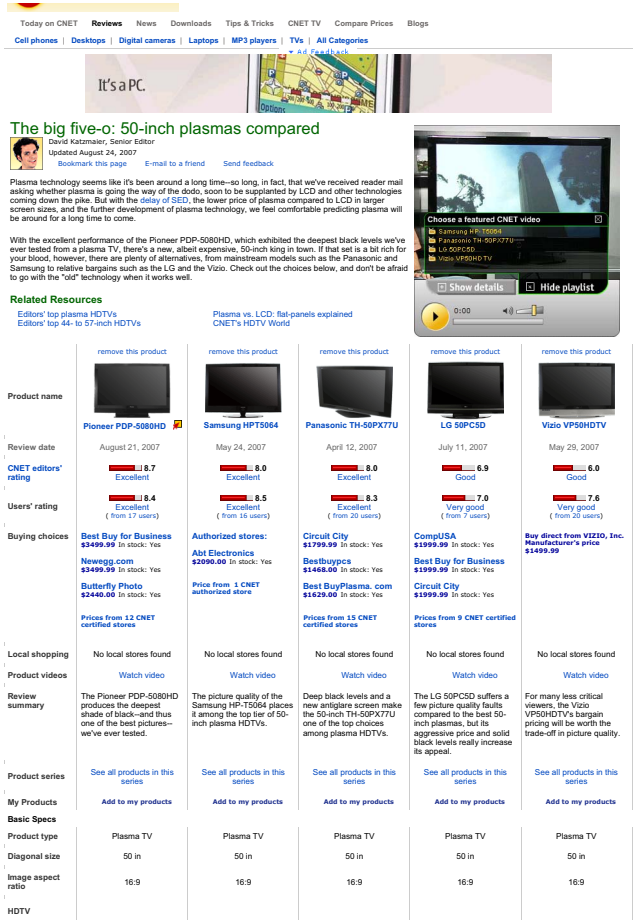
- Por qué: Uno debe tener un objetivo a la hora de recoger unos datos y luego analizarlos. Unos mismos datos pueden soportar diferentes interpretaciones y resultados, así que es necesario querer algo o si no el análisis de datos se convierte en interminable

ACTIVIDADES

EJERCICIO 1.5.1 Identifica en el ejemplo de las bicicletas qué tipo de variables aparecen

EJERCICIO 1.5.2 Qué objetivo crees que persigue el que realizó la investigación

EJERCICIO 1.5.3 En la siguiente tabla identifica qué son los casos, qué variables se utilizan y de qué tipo son las variables que aparecen.



The big five-o: 50-inch plasmas compared

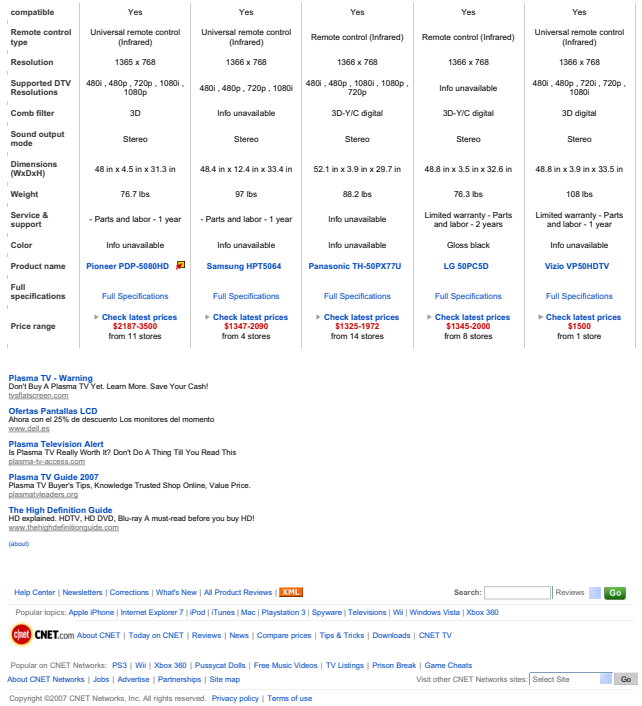
David Katzmaier, Senior Editor
Updated August 24, 2007

Plasma technology seems like it's been around a long time--so long, in fact, that we've received reader mail asking whether plasma is going the way of the dodo, soon to be supplanted by LCD and other technologies coming down the pike. But with the delay of SED, the lower price of plasma compared to LCD in larger screen sizes, and the further development of plasma technology, we feel comfortable predicting plasma will be around for a long time to come.

With the excellent performance of the Pioneer PDP-5080HD, which exhibited the deepest black levels we've ever tested from a plasma TV, there's a new, albeit expensive, 50-inch king in town. If that set is a bit rich for your blood, however, there are plenty of alternatives, from mainstream models such as the Panasonic and Samsung to relative bargains such as the LG and the Vizio. Check out the choices below, and don't be afraid to go with the "old" technology when it works well.

Related Resources

	Pioneer PDP-5080HD	Samsung HPT5064	Panasonic TH-50PX77U	LG 50PC5D	Vizio VP50HDTV
Review date	August 21, 2007	May 24, 2007	April 12, 2007	July 11, 2007	May 29, 2007
CNET editors' rating	8.7 Excellent	8.0 Excellent	8.0 Excellent	6.9 Good	6.0 Good
Users' rating	8.4 Excellent (from 17 users)	8.5 Excellent (from 16 users)	8.3 Excellent (from 20 users)	7.0 Very good (from 7 users)	7.6 Very good (from 20 users)
Buying choices	Best Buy for Business $3499.99 In stock: Yes; Newegg.com $3499.99 In stock: Yes; Butterfly Photo $2440.00 In stock: Yes; Prices from 12 CNET certified stores	Authorized stores: Abt Electronics $2090.00 In stock: Yes; Price from 1 CNET authorized store	Circuit City $1799.99 In stock: Yes; Bestbuypcs $1468.00 In stock: Yes; Best BuyPlasma. com $1629.00 In stock: Yes; Prices from 15 CNET certified stores	CompUSA $1999.99 In stock: Yes; Best Buy for Business $1999.99 In stock: Yes; Circuit City $1999.99 In stock: Yes; Prices from 9 CNET certified stores	Buy direct from VIZIO, Inc. Manufacturer's price $1499.99
Local shopping	No local stores found	No local stores found	No local stores found	No local stores found	No local stores found
Product videos	Watch video	Watch video	Watch video	Watch video	Watch video
Review summary	The Pioneer PDP-5080HD produces the deepest shade of black--and thus one of the best pictures--we've ever tested.	The picture quality of the Samsung HP-T5064 places it among the top tier of 50-inch plasma HDTVs.	Deep black levels and a new antiglare screen make the 50-inch TH-50PX77U one of the top choices among plasma HDTVs.	The LG 50PC5D suffers a few picture quality faults compared to the best 50-inch plasmas, but its aggressive price and solid black levels really increase its appeal.	For many less critical viewers, the Vizio VP50HDTV's bargain pricing will be worth the trade-off in picture quality.
Product series	See all products in this series	See all products in this series	See all products in this series	See all products in this series	See all products in this series
My Products	Add to my products	Add to my products	Add to my products	Add to my products	Add to my products
Basic Specs					
Product type	Plasma TV	Plasma TV	Plasma TV	Plasma TV	Plasma TV
Diagonal size	50 in	50 in	50 in	50 in	50 in
Image aspect ratio	16:9	16:9	16:9	16:9	16:9
HDTV					



compatible	Yes	Yes	Yes	Yes	Yes
Remote control type	Universal remote control (Infrared)	Universal remote control (Infrared)	Remote control (Infrared)	Remote control (Infrared)	Universal remote control (Infrared)
Resolution	1365 x 768	1366 x 768	1366 x 768	1366 x 768	1366 x 768
Supported DTV Resolutions	480i , 480p , 720p , 1080i , 1080p	480i , 480p , 720p , 1080i	480i , 480p , 1080i , 1080p , 720p	Info unavailable	480i , 480p , 720i , 720p , 1080i
Comb filter	3D	Info unavailable	3D-Y/C digital	3D-Y/C digital	3D digital
Sound output mode	Stereo	Stereo	Stereo	Stereo	Stereo
Dimensions (WxDxH)	48 in x 4.5 in x 31.3 in	48.4 in x 12.4 in x 33.4 in	52.1 in x 3.9 in x 29.7 in	48.8 in x 3.5 in x 32.6 in	48.8 in x 3.9 in x 33.5 in
Weight	76.7 lbs	97 lbs	88.2 lbs	76.3 lbs	108 lbs
Service & support	- Parts and labor - 1 year	- Parts and labor - 1 year	Info unavailable	Limited warranty - Parts and labor - 2 years	Limited warranty - Parts and labor - 1 year
Color	Info unavailable	Info unavailable	Info unavailable	Gloss black	Info unavailable
Product name	Pioneer PDP-5080HD	Samsung HPT5064	Panasonic TH-50PX77U	LG 50PC5D	Vizio VP50HDTV
Full specifications	Full Specifications	Full Specifications	Full Specifications	Full Specifications	Full Specifications
Price range	Check latest prices $2187-3500 from 11 stores	Check latest prices $1347-2090 from 4 stores	Check latest prices $1325-1972 from 14 stores	Check latest prices $1345-2000 from 8 stores	Check latest prices $1500 from 1 store

Plasma TV - Warning
Don't Buy A Plasma TV Yet. Learn More. Save Your Cash!
tvsflatscreen.com

Ofertas Pantallas LCD
Ahora con el 25% de descuento Los monitores del momento
www.dell.es

Plasma Television Alert
Is Plasma TV Really Worth It? Don't Do A Thing Till You Read This
plasma-tv-access.com

Plasma TV Guide 2007
Plasma TV Buyer's Tips, Knowledge Trusted Shop Online, Value Price.
plasmatvleaders.org

The High Definition Guide
HD explained. HDTV, HD DVD, Blu-ray A must-read before you buy HD!
www.thehighdefinitionguide.com

(about)

Help Center | Newsletters | Corrections | What's New | All Product Reviews | XML Search: Reviews Go

Popular topics: Apple iPhone | Internet Explorer 7 | iPod | iTunes | Mac | Playstation 3 | Spyware | Televisions | Wii | Windows Vista | Xbox 360

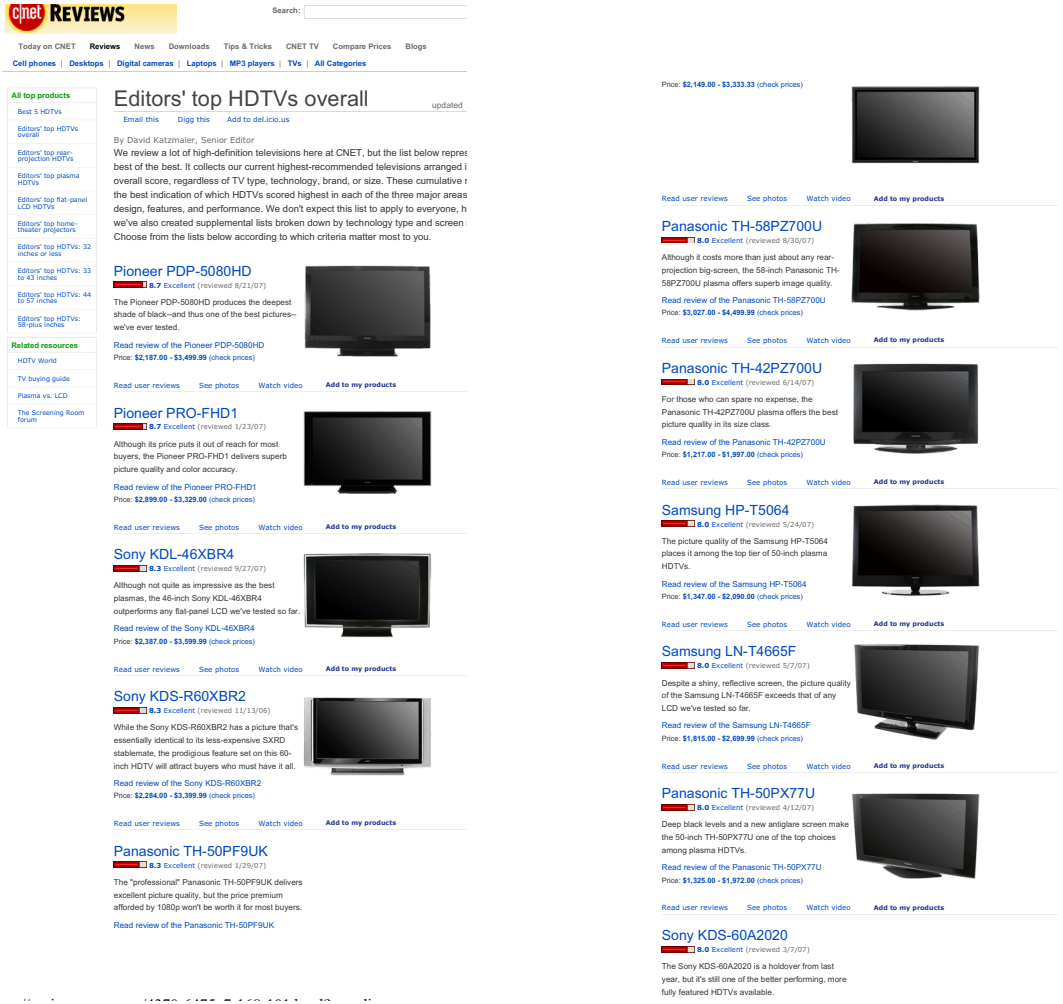
CNET.com About CNET | Today on CNET | Reviews | News | Compare prices | Tips & Tricks | Downloads | CNET TV

Popular on CNET Networks: PS3 | Wii | Xbox 360 | Pussycat Dolls | Free Music Videos | TV Listings | Prison Break | Game Cheats

About CNET Networks | Jobs | Advertise | Partnerships | Site map Visit other CNET Networks sites: Select Site Go

http://reviews.cnet.com/4321-6482_7-6575571.html?tag=prmo1 01/10/2007

EJERCICIO 1.5.4 Identifica casos, variables, etc.



cnet REVIEWS

Search:

Today on CNET | Reviews | News | Downloads | Tips & Tricks | CNET TV | Compare Prices | Blogs

Cell phones | Desktops | Digital cameras | Laptops | MP3 players | TVs | All Categories

All top products
- Best 5 HDTVs
- Editors' top HDTVs overall
- Editors' top rear-projection HDTVs
- Editors' top plasma HDTVs
- Editors' top flat-panel LCD HDTVs
- Editors' top home-theater projectors
- Editors' top HDTVs: 32 inches or less
- Editors' top HDTVs: 33 to 43 inches
- Editors' top HDTVs: 44 to 57 inches
- Editors' top HDTVs: 58-plus inches

Related resources
- HDTV World
- TV buying guide
- Plasma vs. LCD
- The Screening Room forum

Editors' top HDTVs overall

updated

Email this | Digg this | Add to del.icio.us

By David Katzmaier, Senior Editor

We review a lot of high-definition televisions here at CNET, but the list below represe
best of the best. It collects our current highest-recommended televisions arranged i
overall score, regardless of TV type, technology, brand, or size. These cumulative r
the best indication of which HDTVs scored highest in each of the three major areas
design, features, and performance. We don't expect this list to apply to everyone, h
we've also created supplemental lists broken down by technology type and screen
Choose from the lists below according to which criteria matter most to you.

Pioneer PDP-5080HD

8.7 Excellent (reviewed 8/21/07)

The Pioneer PDP-5080HD produces the deepest shade of black--and thus one of the best pictures-- we've ever tested.

Read review of the Pioneer PDP-5080HD

Price: $2,187.00 - $3,499.99 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Pioneer PRO-FHD1

8.7 Excellent (reviewed 1/23/07)

Although its price puts it out of reach for most buyers, the Pioneer PRO-FHD1 delivers superb picture quality and color accuracy.

Read review of the Pioneer PRO-FHD1

Price: $2,899.00 - $3,329.00 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Sony KDL-46XBR4

8.3 Excellent (reviewed 9/27/07)

Although not quite as impressive as the best plasmas, the 46-inch Sony KDL-46XBR4 outperforms any flat-panel LCD we've tested so far.

Read review of the Sony KDL-46XBR4

Price: $2,387.00 - $3,599.99 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Sony KDS-R60XBR2

8.3 Excellent (reviewed 11/13/06)

While the Sony KDS-R60XBR2 has a picture that's essentially identical to its less-expensive SXRD stablemate, the prodigious feature set on this 60-inch HDTV will attract buyers who must have it all.

Read review of the Sony KDS-R60XBR2

Price: $2,284.00 - $3,399.99 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Panasonic TH-50PF9UK

8.3 Excellent (reviewed 1/29/07)

The "professional" Panasonic TH-50PF9UK delivers excellent picture quality, but the price premium afforded by 1080p won't be worth it for most buyers.

Read review of the Panasonic TH-50PF9UK

Price: $2,149.00 - $3,333.33 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Panasonic TH-58PZ700U

8.0 Excellent (reviewed 8/30/07)

Although it costs more than just about any rear-projection big-screen, the 58-inch Panasonic TH-58PZ700U plasma offers superb image quality.

Read review of the Panasonic TH-58PZ700U

Price: $3,027.00 - $4,499.99 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Panasonic TH-42PZ700U

8.0 Excellent (reviewed 6/14/07)

For those who can spare no expense, the Panasonic TH-42PZ700U plasma offers the best picture quality in its size class.

Read review of the Panasonic TH-42PZ700U

Price: $1,217.00 - $1,997.00 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Samsung HP-T5064

8.0 Excellent (reviewed 5/24/07)

The picture quality of the Samsung HP-T5064 places it among the top tier of 50-inch plasma HDTVs.

Read review of the Samsung HP-T5064

Price: $1,347.00 - $2,090.00 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Samsung LN-T4665F

8.0 Excellent (reviewed 5/7/07)

Despite a shiny, reflective screen, the picture quality of the Samsung LN-T4665F exceeds that of any LCD we've tested so far.

Read review of the Samsung LN-T4665F

Price: $1,815.00 - $2,699.99 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Panasonic TH-50PX77U

8.0 Excellent (reviewed 4/12/07)

Deep black levels and a new antiglare screen make the 50-inch TH-50PX77U one of the top choices among plasma HDTVs.

Read review of the Panasonic TH-50PX77U

Price: $1,325.00 - $1,972.00 (check prices)

Read user reviews | See photos | Watch video | Add to my products

Sony KDS-60A2020

8.0 Excellent (reviewed 3/7/07)

The Sony KDS-60A2020 is a holdover from last year, but it's still one of the better performing, more fully featured HDTVs available.

Parte II
Representando y Describiendo Datos Categóricos

2.1.Ejemplo

Sexo y Divorcio

- En un estudio del año 1979 se recogieron unos datos sobre cuatro variables a 1036 personas (aprox. la mitad había solicitado el divorcio). Las variables eran:
  - Haber tenido relaciones extramaritales
  - Haber tenido relaciones premaritales
  - El género
  - Si habían solicitado el divorcio o seguían casadas

- Una forma de representar estos datos acorde con el estilo que hemos visto en el tema anterior es el siguiente (no están puestos los 1036, sólo hay 9):

	Premarital_	Extramarital_	Married_	Gender_
1	N	N	Divorced	Male
2	N	N	Divorced	Male
3	N	N	Divorced	Male
4	N	N	Divorced	Male
5	N	N	Divorced	Male
6	N	N	Divorced	Male
7	N	N	Divorced	Male
8	N	N	Divorced	Male
9	N	N	Divorced	Male

- Ahora bien, cuando se trabaja con datos categóricos, es bastante habitual hacer un recuento y presentar los datos organizados de una manera diferente. Hay muchas posibilidades:

				Extramarital_			
				N		Y	
				Premarital_		Premarital_	
				N	Y	N	Y
				Recuento	Recuento	Recuento	Recuento
Gender_	Fema	Married_	Divorced	214	54	36	17
			Married	322	25	4	4
	Male	Married_	Divorced	68	60	17	28
			Married	130	42	4	11

				Extramarital_			
				N		Y	
				Gender_		Gender_	
				Fema	Male	Fema	Male
				Recuento	Recuento	Recuento	Recuento
Premarital_	N	Married_	Divorced	214	68	36	17
			Married	322	130	4	4
	Y	Married_	Divorced	54	60	17	28
			Married	25	42	4	11

		Married_							
		Divorced				Married			
		Extramarital_				Extramarital_			
		N		Y		N		Y	
		Gender_		Gender_		Gender_		Gender_	
		Fema	Male	Fema	Male	Fema	Male	Fema	Male
		Recuento	Recuento	Recuento	Recuento	Recuento	Recuento	Recuento	Recuento
Premarital_	N	214	68	36	17	322	130	4	4
	Y	54	60	17	28	25	42	4	11

- La mejor forma de presentar una tabla de frecuencias empieza planteando cual es la variable que nos interesa explicar (usualmente llamada la dependiente)
  - En el caso de los datos de sexo, lo que interesa es ver qué conductas o variables llevan a que la gente se divorcie más

– Una vez decidido cuál es la variable que más nos interesa ponemos las otras en las filas y la interesante en columnas de esta manera:

						Married_	
						Divorced	Married
						Recuento	Recuento
Gender_	Fema	Extramarital_	N	Premarital_	N	214	322
					Y	54	25
			Y	Premarital_	N	36	4
					Y	17	4
	Male	Extramarital_	N	Premarital_	N	68	130
					Y	60	42
			Y	Premarital_	N	17	4
					Y	28	11

- Si además calculamos porcentajes sobre la última variable entonces podemos hacer ya algunas observaciones interesantes:

						Married_	
						Divorced	Married
						Recuento	Recuento
Gender_	Fema	Extramarital_	N	Premarital_	N	40	60
					Y	68	32
			Y	Premarital_	N	90	10
					Y	81	19
	Male	Extramarital_	N	Premarital_	N	34	66
					Y	59	41
			Y	Premarital_	N	81	19
					Y	72	28

  - Tener en cuenta que había 494 (47.7%) divorciados y 542 (52.3%) casados
  - Cualquier desviación del 47.7%-52.3% indicaría que hay una abundancia/escasez en las categorías de divorciado/casado

2.2.Representando y evaluando datos categóricos

- En el ejemplo anterior hemos visto unos datos con cuatro variables categóricas
- En la práctica, los datos categóricos se suelen trabajar viendo:
  - una variable cada vez (por ejemplo, Género o divorciado casado)
  - dos variables cada vez mediante “cruces” o “tablas” (por ejemplo, Género con divorciado/casado, o Género con Relaciones Prematrimoniales)
- Ver más de dos variables a la vez resulta raro pero es posible (como hemos visto anteriormente)

2.3.Representación gráfica de una variable categórica

- Para representar una variable categórica se puede usar:
  - Un diagrama de pastel
  - Un diagrama de barras
  - Un gráfico supercalifragilístico que es básicamente cualquiera de los otros dos pero con un montón de tinta superficial
- Veremos estos gráficos a continuación

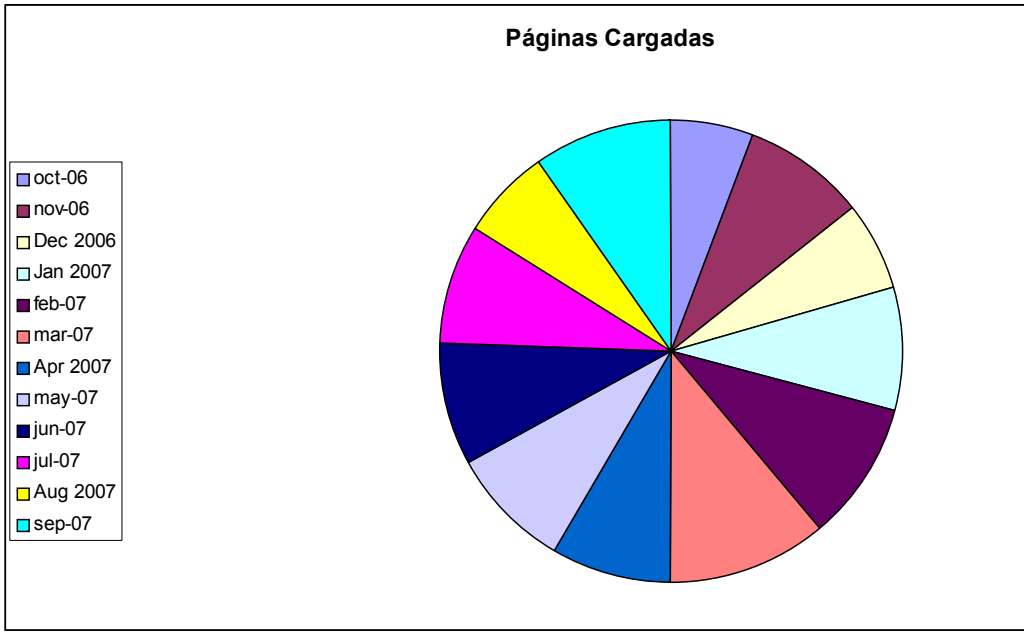
2.4.Diagramas de Pastel

Ejemplo

El siguiente ejemplo corresponde a los datos de visitas durante los últimos 12 meses a una página web sobre un libro. El objetivo de hacer estas representaciones es evaluar qué meses se reciben más visitas, cuales son los máximos y los mínimos de visitas, etc.

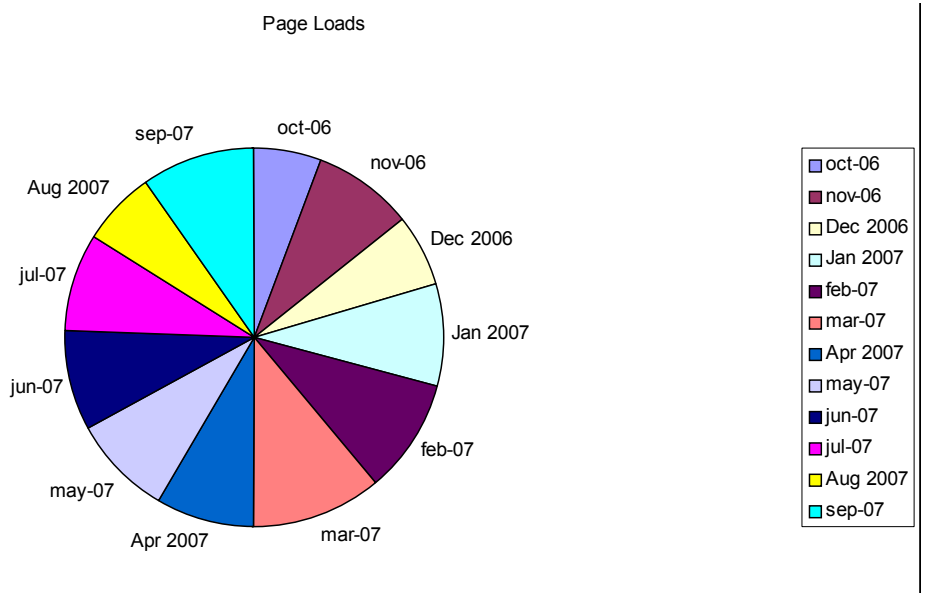
- Los diagramas de pastel son muy populares para datos de este tipo aunque tiene algunos defectos que veremos a continuación

- En primer lugar, veamos un mal ejemplo(Excel por defecto):



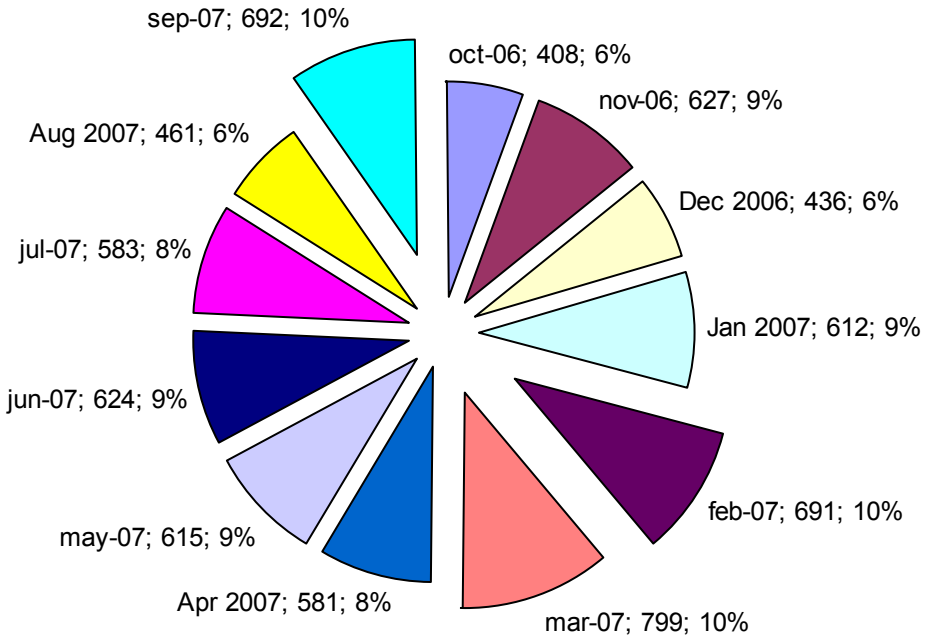
  - Las etiquetas están a la izquierda y hay que ir mirando una por una->Aburrido

- Mirando las opciones de Excel podemos mejorarlo algo



  - En este tipo de gráficos hay que poner normalmente la etiqueta junto a cada porción del gráfico
  - ¿Qué mes tiene más visitas? ¿o menos? Todos los trozos parecen iguales! ¿Y los porcentajes?

- Ajajá! Esto está mucho mejor!

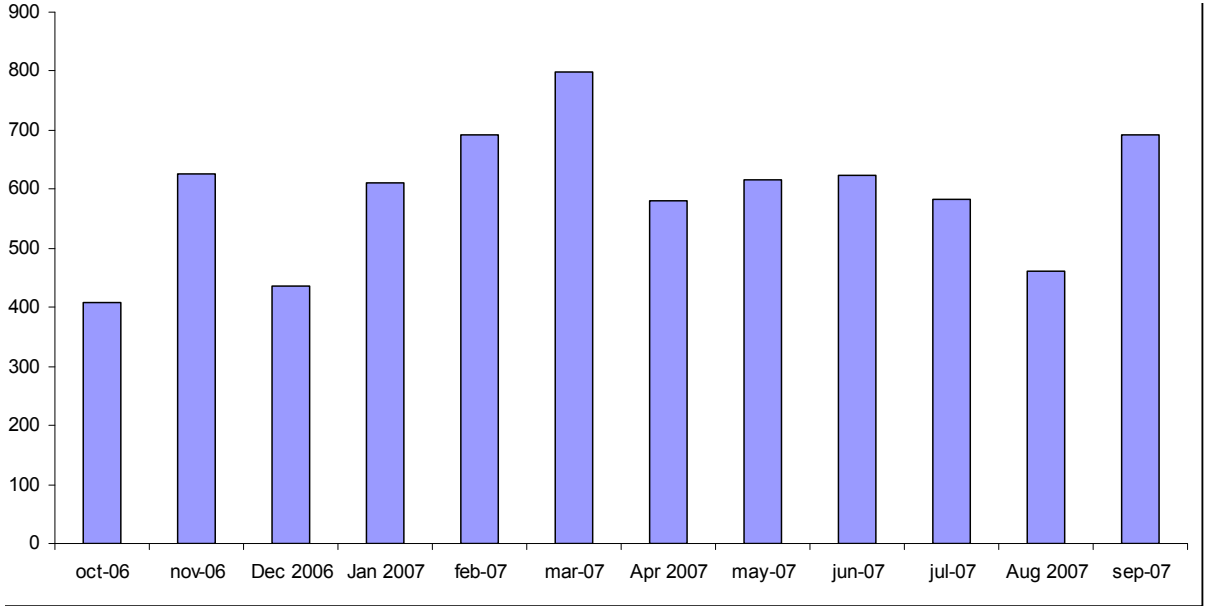


  - El gráfico ahora tiene toda la información pero aparece muy recargado
  - Sacar trozos del pastel para recalcar un dato es interesante, pero si hay varios queda mal

- En resumen:
  - Los diagramas de pastel se usan mucho porque se entienden bien, y quedan bien en el papel
  - No obstante, hay que tener cuidado y sentido estético para transmitir lo que se quiere transmitir sin embrollarlo todo
  - Tener cuidado con el color. Si luego haceis la impresión en blanco y negro (lo cual es bastante recomendable) los gráficos de pastel quedan horrorosos

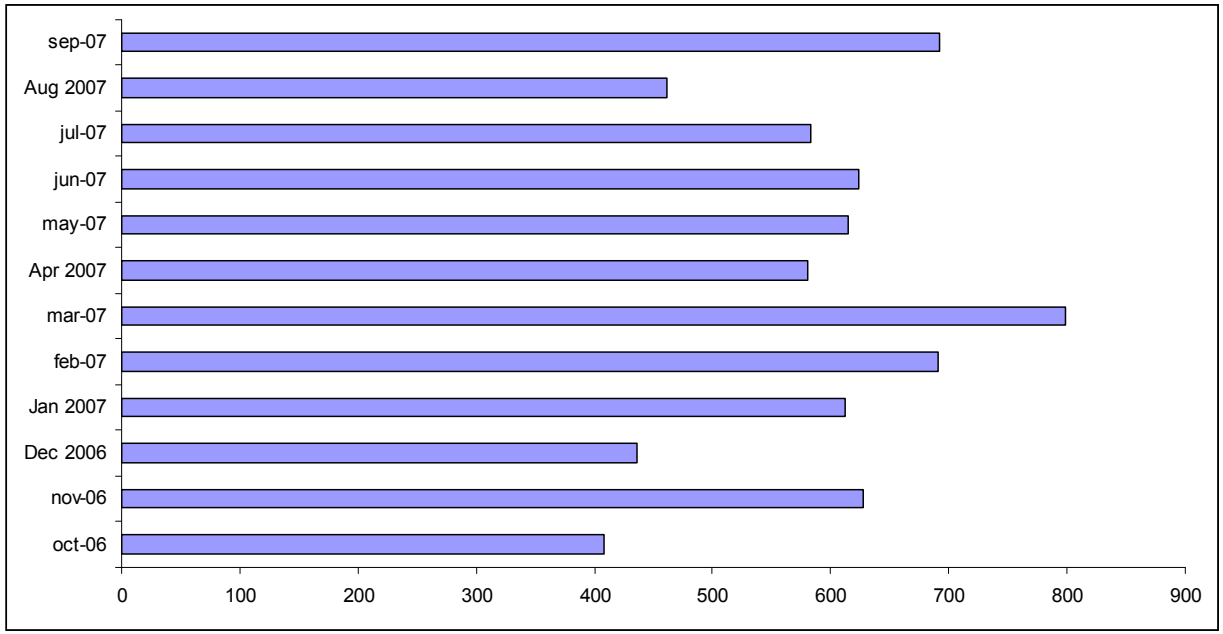
2.5.Diagramas de barras

- La idea es poner los recuentos en función de la altura de la barra.



  - Busquemos ahora máximos y mínimos ¿Es fácil?

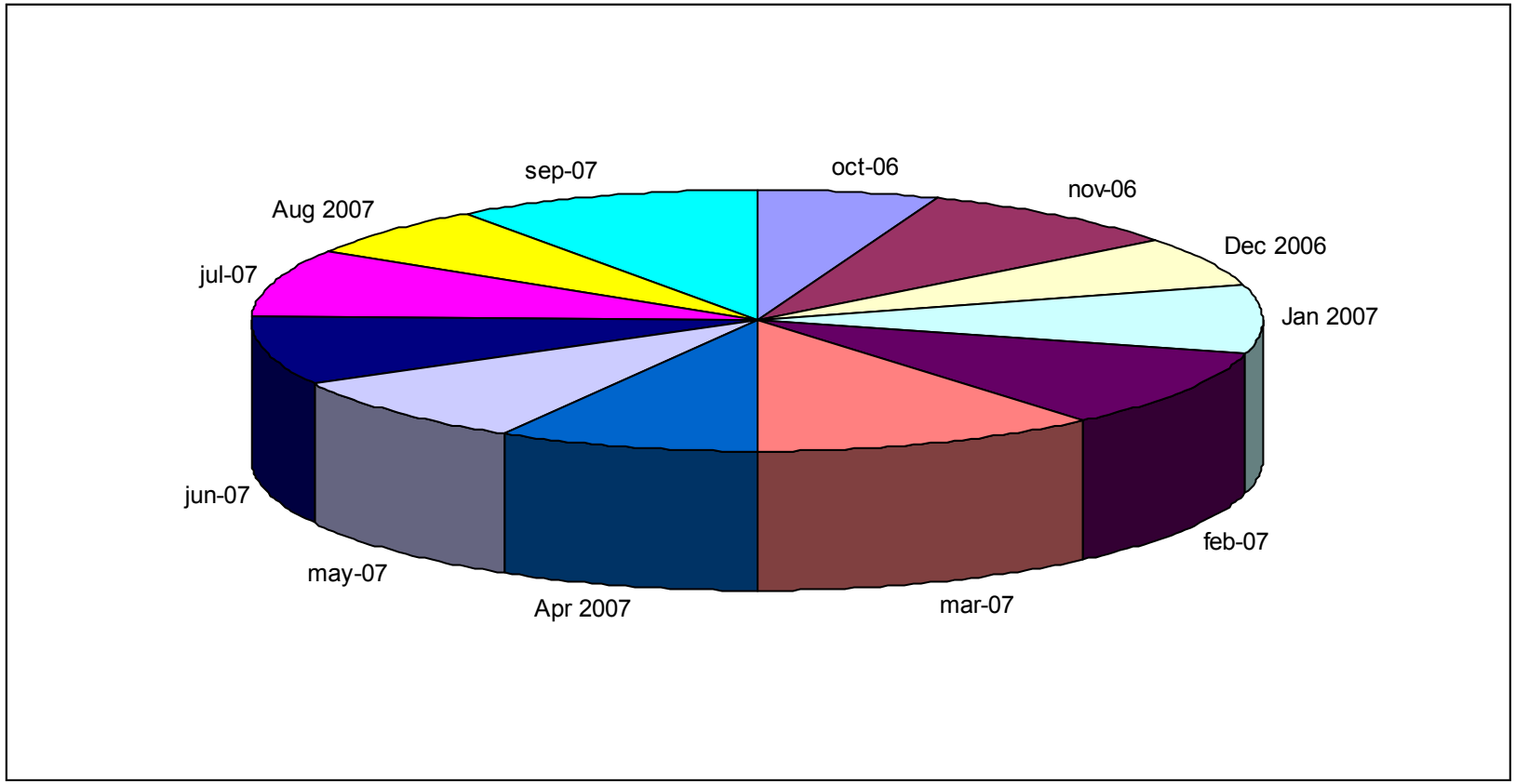
- También se puede poner de lado según la forma de la página



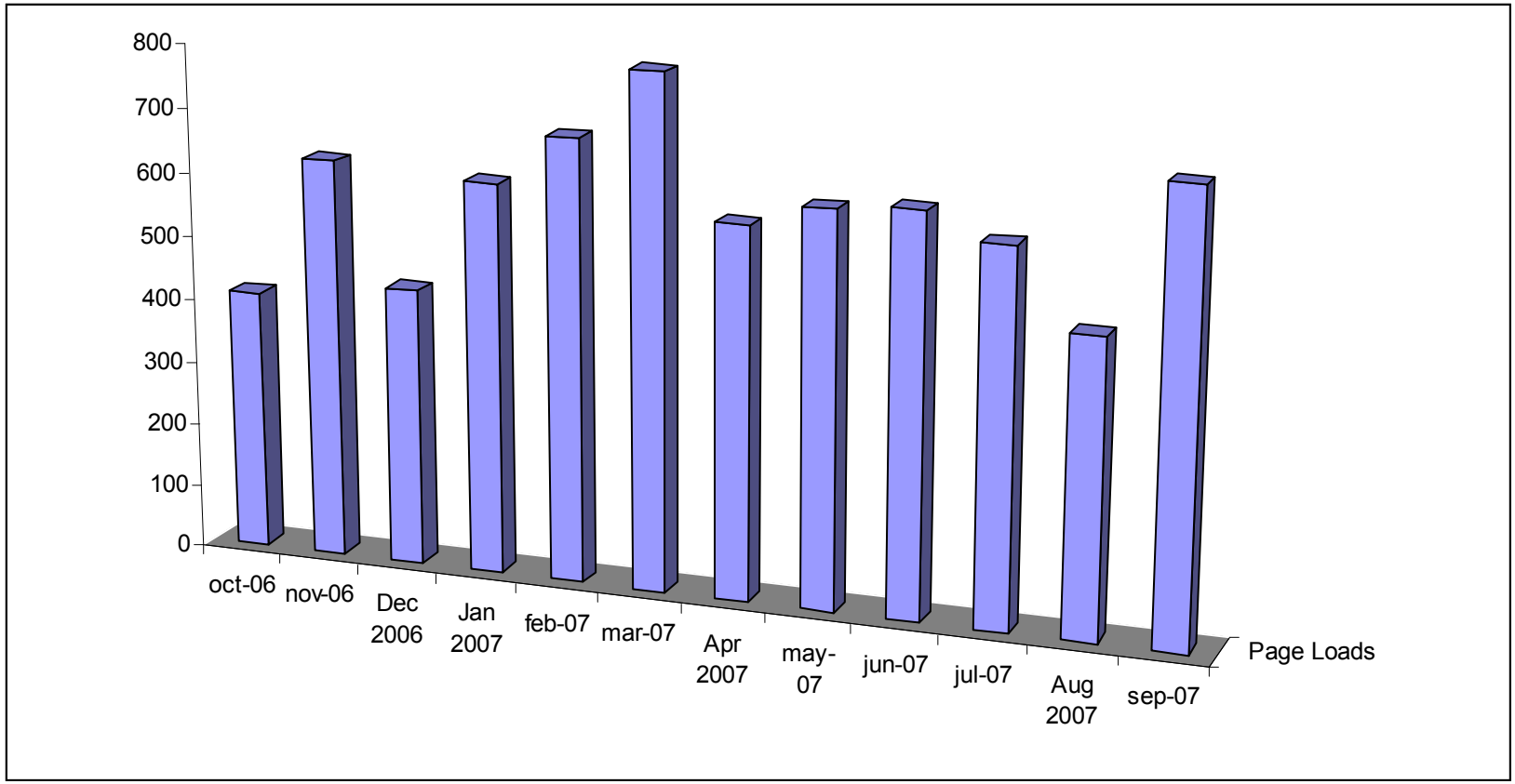
- Las barras pueden estar ordenadas siguiendo cualquier criterio
  - Se pueden ordenar en función del valor que tengan (las más grandes al principio y luego en sentido decreciente)
  - En este caso están ordenadas según la serie temporal
  - Por orden alfabético puede ser aceptable si no hay nada mejor

2.6.Ideas supercalifragilísticas

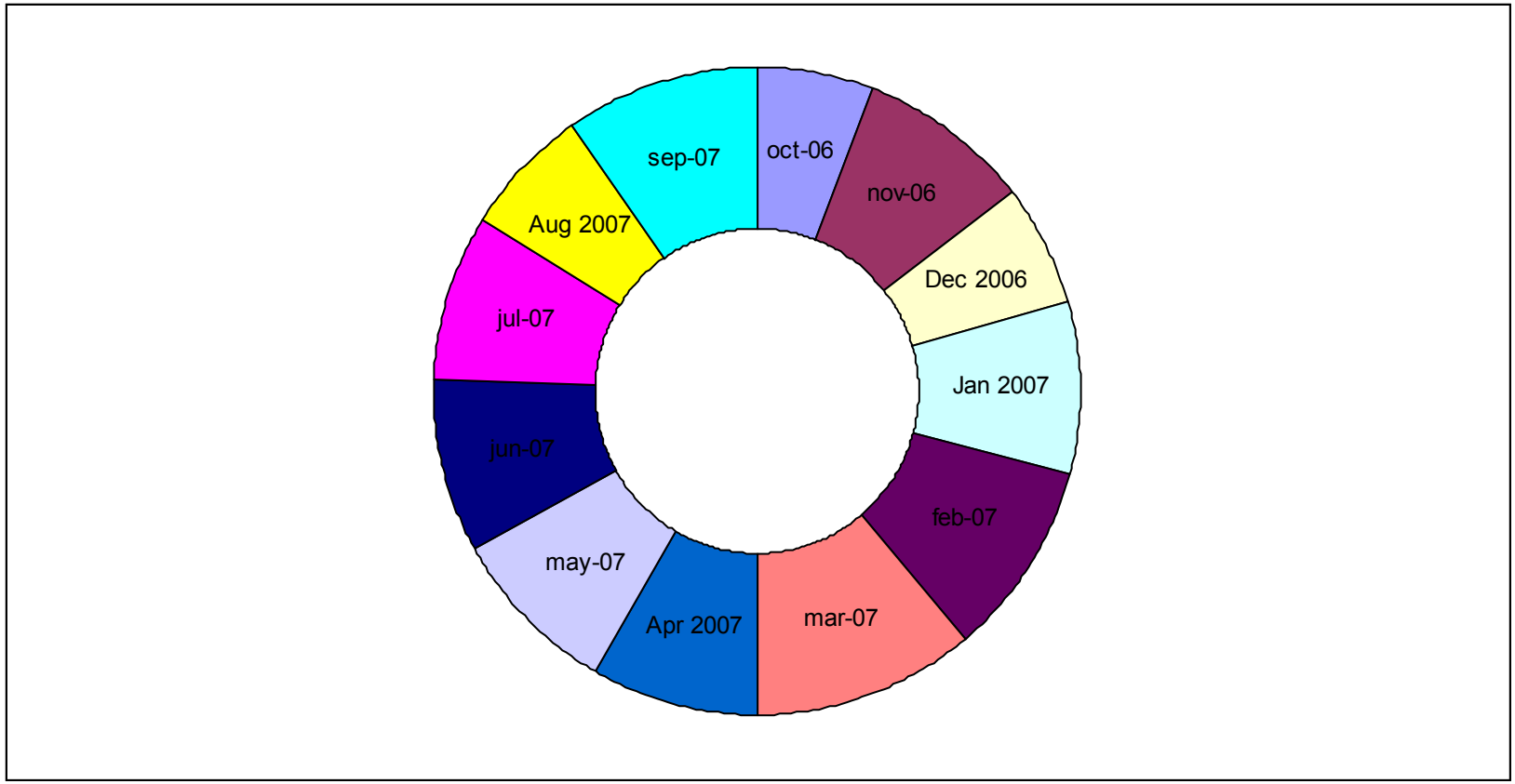
- Excel nos ofrece unas cuantas de este tipo. Por ejemplo



- O también:


800
700
600
500
400
300
200
100
0
oct-06
nov-06
Dec 2006
Jan 2007
feb-07
mar-07
Apr 2007
may-07
jun-07
jul-07
Aug 2007
sep-07
Page Loads

- Esto es interminable



2.7.Resumen sobre representación de una variable categórica

- Los diagramas de barras suelen ser la opción más razonable
- A veces es mejor poner una tabla de datos con porcentajes
- Los diagramas de pastel pueden ser aceptables si se cuidan los detalles
- Los diagramas de barras son la opción más simple y a menudo la mejor
- Huir de los gráficos supercalifragilísticos (o meteros a diseñador gráfico)

2.8.Trabajando con dos variables categóricas a la vez

- En muchos estudios se suelen plantear los análisis teniendo en cuenta dos variables a la vez. Para el ejemplo de Sexo podemos querer ver:
  - Qué genero se divorcia más.
  - Qué género tiene más relaciones prematrimoniales, o extramatrimoniales.
  - Qué relación hay entre tener relaciones Prematrimoniales o Extramatrimoniales y el divorcio.

- Este tipo de preguntas se suele mostrar como una tabla de frecuencias cruzadas. Por ejemplo:

GENDER	Divorced	Married	Row Sums
Male	173	187	360
Female	321	355	676
Column Sums	494	542	1036

  - En estas tablas se suelen poner las sumas por filas y por columnas (Vemos que hay más mujeres que hombres y más casados que solteros)

- Las tablas de frecuencias cruzadas pueden ser un poco engañosas de interpretar:
  - Por ejemplo, una interpretación muuuuuy ingénua de la tabla anterior se fijaría en que hay 355 mujeres casadas y podría concluir que las mujeres tienden sobre todo a no divorciarse
  - Esa interpretación no tiene sentido ya que para valorar una frecuencia dentro de la tabla hay que tener en cuenta los totales de la tabla.
  - Así, para valorar si 355 es un valor alto para las mujeres casadas hay que verlo en relación con el número total de mujeres que hay en los datos. En este caso, 355 de 676 no parece demasiado.

- Una forma de ver un valor en relación con otro es calculando una división por el total por fila
  - Por ejemplo, 355 mujeres casadas dividido por el total de mujeres que es 676 es 0.5251. Multiplicando ese valor por 100 tenemos el porcentaje (52.51%) de casadas dado que son mujeres. Haciéndolo para todas las casillas tenemos la tabla de abajo

GENDER	Divorced	Married	Row Sums
Male	48.06	51.94	100.00
Female	47.49	52.51	100.00

  - Esto se llama porcentajes por filas

- No obstante, también se puede calcular en relación con los totales por columnas.
  - Si dividimos 355 por el número de personas casadas que es 542 y multiplicamos por 100 tenemos 65.5%. Este valor es el porcentaje de mujeres dado que se está casado

	MARRIED?	
GENDER	Divorced	Married
Male	35.02	34.50
Female	64.98	65.50
Column Sums	100.00	100.00

  - Esto se llama porcentaje por columna

- Llamarlos porcentajes por columna o por fila es un poco arbitrario ya que podemos intercambiar la variable que está en columnas por la que está en filas. Por ejemplo:
  - Intercambiando filas y columnas y calculando porcentaje por filas tenemos

% de Gender_

		Married_		
		Divorced	Married	Total
Gender_	Fema	47.5%	52.5%	100.0%
	Male	48.1%	51.9%	100.0%
Total		47.7%	52.3%	100.0%

% de Married_

		Gender_		
		Fema	Male	Total
Married_	Divorced	65.0%	35.0%	100.0%
	Married	65.5%	34.5%	100.0%
Total		65.3%	34.7%	100.0%

- ¿Hay una regla general para hacer este tipo de porcentajes en tablas?
  - En principio, no existe una regla concreta acerca de como hay que hacer esto pero yo recomiendo poner la variable explicadora (ya sabeis, la independiente) en las filas, y la explicada (la dependiente) en las columnas
  - Luego se calculan los porcentajes por filas.
  - En nuestro caso, si queremos explicar el divorcio en función del género, recomiendo hacer esta:

% de Gender_

		Married_		
		Divorced	Married	Total
Gender_	Fema	47.5%	52.5%	100.0%
	Male	48.1%	51.9%	100.0%
Total		47.7%	52.3%	100.0%

- Todo lo de antes está muy bien, pero ¿cómo se interpreta?
  - Si se han seguido las reglas de antes siempre podemos decir: El porcentaje de los <aquí categoría de fila> que son/están <aquí categoría de las columna> es <aquí porcentaje>
  - Por ejemplo, en la tabla de abajo, el porcentaje de las mujeres que están divorciadas es el 47.5%

% de Gender_

		Married_		
		Divorced	Married	Total
Gender_	Fema	47.5%	52.5%	100.0%
	Male	48.1%	51.9%	100.0%
Total		47.7%	52.3%	100.0%

  - Fijaros que esto no es correcto, el porcentaje de los divorciados que son mujeres es el 47.5% (el valor correcto es 65%)

- Fantástico, pero ¿cuándo puedo sacar una conclusión interesante de estas tablas?
  - La forma de ver si un valor es llamativo es compararlo con los porcentajes totales en las filas

% de Gender_

		Married_		
		Divorced	Married	Total
Gender_	Fema	47.5%	52.5%	100.0%
	Male	48.1%	51.9%	100.0%
Total		47.7%	52.3%	100.0%

  - Los porcentajes por filas de divorciados y casados es 47.7% y 52.3%. Vemos que esos porcentajes atendidendo a si son hombres o mujeres son muy similares así que concluimos que ser hombre o mujer no tiene mucho efecto sobre el estar casado o no.

- ¿Hay más maneras de sacar los porcentajes de las tablas?
  - Una última posibilidad es sacar los porcentajes con respecto al total de la tabla y no con respecto a las filas o las columnas. Por ejemplo:

Tabla de contingencia Gender_ * Married_

% del total

		Married_		
		Divorced	Married	Total
Gender_	Fema	31.0%	34.3%	65.3%
	Male	16.7%	18.1%	34.7%
Total		47.7%	52.3%	100.0%

  - Esta tabla nos permite tener una idea de la importancia relativa de cada celda.

ACTIVIDADES

EJERCICIO 2.8.1 Los datos de supervivencia del hundimiento del Titanic se usan en muchas ocasiones como un ejemplo de análisis de datos categóricos. En este caso, estudiaremos la relación entre el tipo de pasajero (de primera clase, de segunda, tercera o miembro de la tripulación) y si sobrevivieron cuando se hundió el barco. A continuación puedes ver una tabla de este análisis. A partir de esta tabla, ¿qué tipo de pasajero dirias que corrió mejor suerte?

Recuento

		Survive_		
		Died	Lived	Total
Class_	1st	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Cre	673	212	885
Total		1490	711	2201

EJERCICIO 2.8.2 ¿Y a partir de esta tabla?

Tabla de contingencia Class_ * Survive_

% de Class_

		Survive_		
		Died	Lived	Total
Class_	1st	37.5%	62.5%	100.0%
	2nd	58.6%	41.4%	100.0%
	3rd	74.8%	25.2%	100.0%
	Cre	76.0%	24.0%	100.0%
Total		67.7%	32.3%	100.0%

EJERCICIO 2.8.3 En líneas generales, ¿dirias que hay una relación entre el tipo de pasajero y sobrevivir o no? ¿Qué tipo de pasajero fue el que peor lo pasó?

EJERCICIO 2.8.4 Todos hemos oido la frase de “las mujeres y los niños primero”. ¿Se cumplió en el Titanic según esta tabla?

Tabla de contingencia Gender_ * Survive_

Recuento

		Survive_		
		Died	Lived	Total
Gender_	Fema	126	344	470
	Male	1364	367	1731
Total		1490	711	2201

EJERCICIO 2.8.5 ¿Esta tabla tiene algún fallo teniendo en cuenta lo que os he enseñado?

Tabla de contingencia Survive_ * Class_

Recuento

		Class_				
		1st	2nd	3rd	Cre	Total
Survive_	Died	122	167	528	673	1490
	Lived	203	118	178	212	711
Total		325	285	706	885	2201

EJERCICIO 2.8.6 ¿Qué podrias decir sobre “las mujeres y los niños primero” a partir de esta tabla?

Tabla de contingencia Age_ * Survive_

% de Survive_

		Survive_		
		Died	Lived	Total
Age_	Adult	96.5%	92.0%	95.0%
	Child	3.5%	8.0%	5.0%
Total		100.0%	100.0%	100.0%

EJERCICIO 2.8.7 ¿Y con esta?

Tabla de contingencia Age_ * Survive_

% de Age_

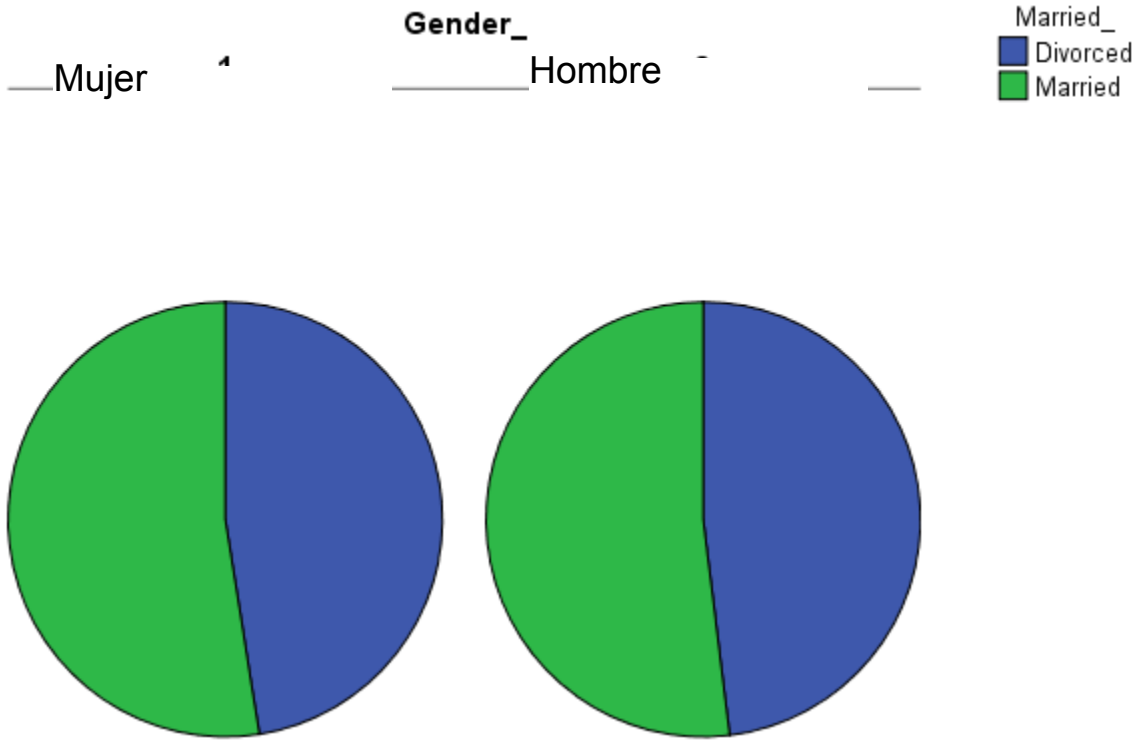
		Survive_		
		Died	Lived	Total
Age_	Adult	68.7%	31.3%	100.0%
	Child	47.7%	52.3%	100.0%
Total		67.7%	32.3%	100.0%

2.9.Representaciones gráficas para tablas de contingencia

- Las tablas de contingencia que hemos visto en la sección anterior se pueden convertir en gráficos estadísticos. Las opciones habituales son:
  - Usar varios gráficos univariados (de pastel o de barras)
  - Usar gráficos de barras partidos
  - Usar diagramas de mosaico
- Estos gráficos aportan más interés a un texto aunque a veces hay que tener cuidado al usarlos ya que pueden ser excesivos para el propósito

2.10.Varios diagramas de pastel o de barras

- Aquí tenemos un diagrama para ver el status marital en función del género



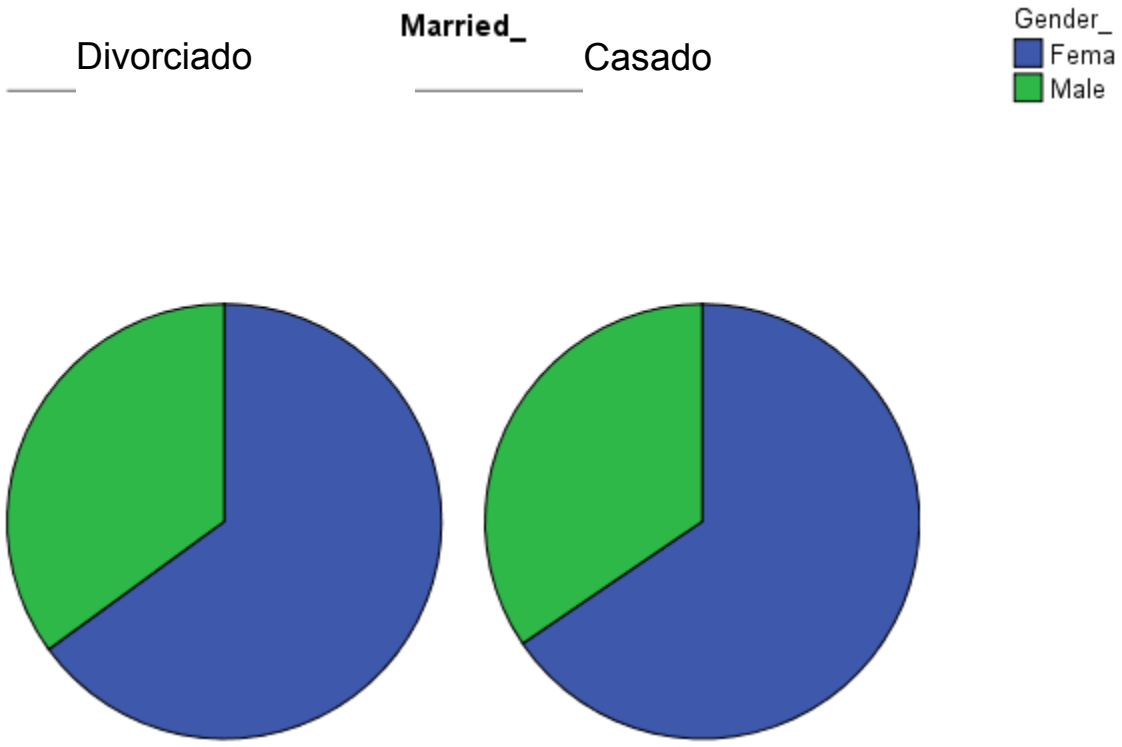
– Fijaros que este gráfico es equivalente a la tabla siguiente de porcentajes por filas

Tabla de contingencia Gender_ * Married_

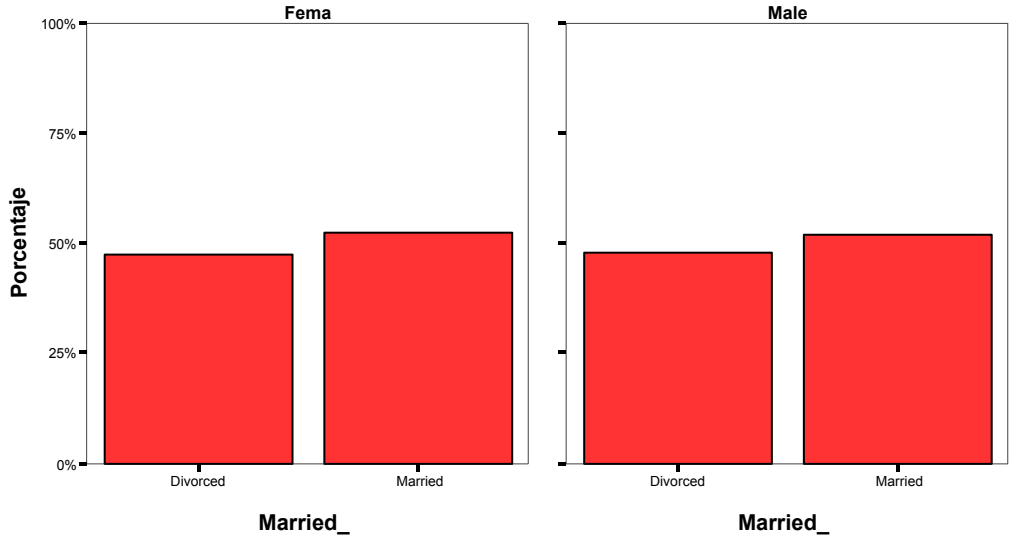
% de Gender_

		Married_		
		Divorced	Married	Total
Gender_	Fema	47.5%	52.5%	100.0%
	Male	48.1%	51.9%	100.0%
Total		47.7%	52.3%	100.0%

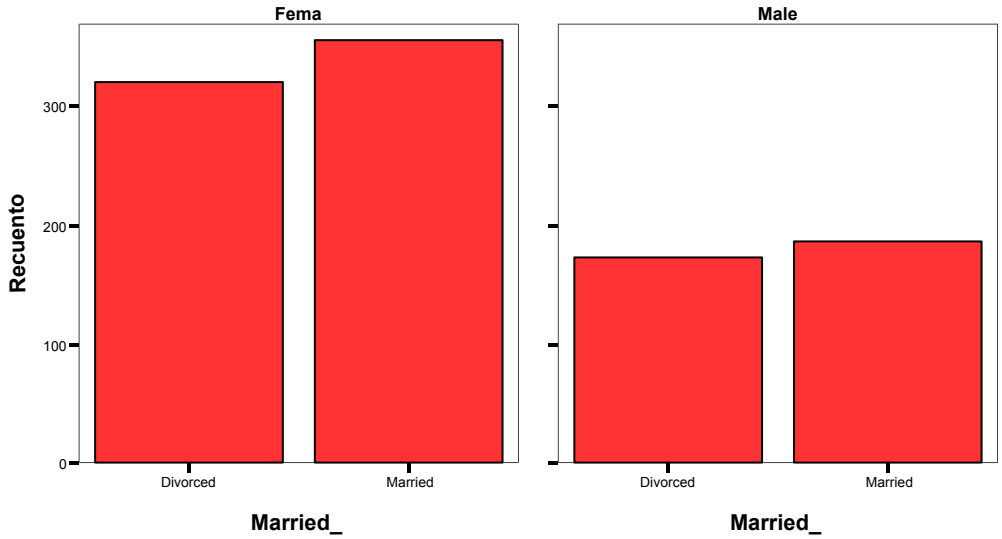
- En cambio, este otro gráfico sería el equivalente a los porcentajes por columnas



- Los mismos gráficos los podemos ver como diagramas de barras. Fijaros que estos gráficos están hechos sobre porcentajes, no sobre los valores absolutos. Eso los hace equivalentes a los gráficos de sectores.

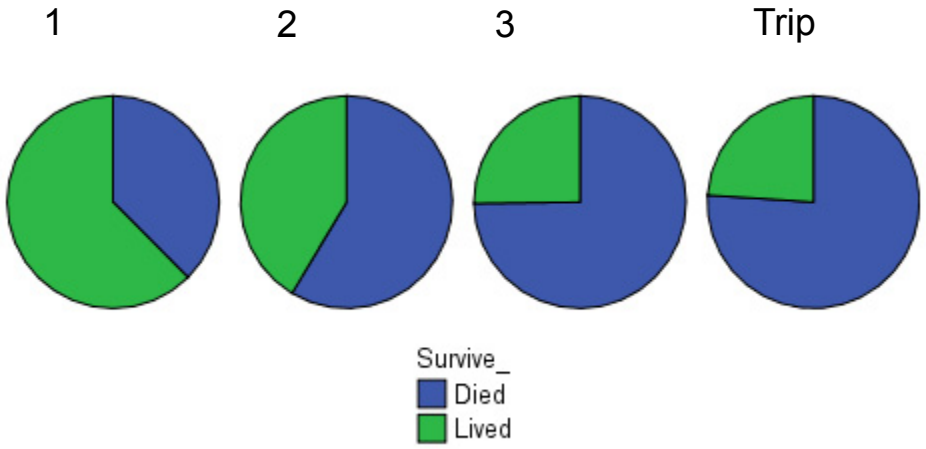


- Esta es la alternativa usando valores absolutos. En realidad la diferencia está en que el gráfico no está escalado al total

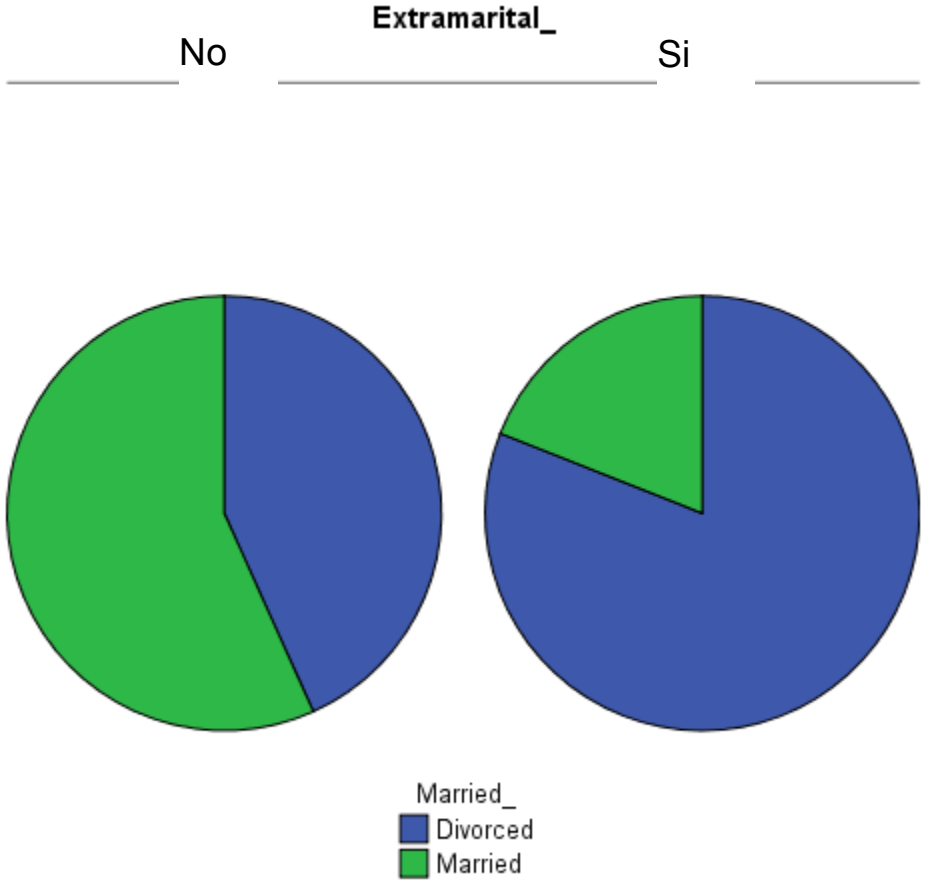


ACTIVIDADES

EJERCICIO 2.10.1 Interpreta el siguiente gráfico de la mortalidad en el titanic



EJERCICIO 2.10.2 ¿Tener relaciones extramaritales tiene influencia sobre divorciarse o no?



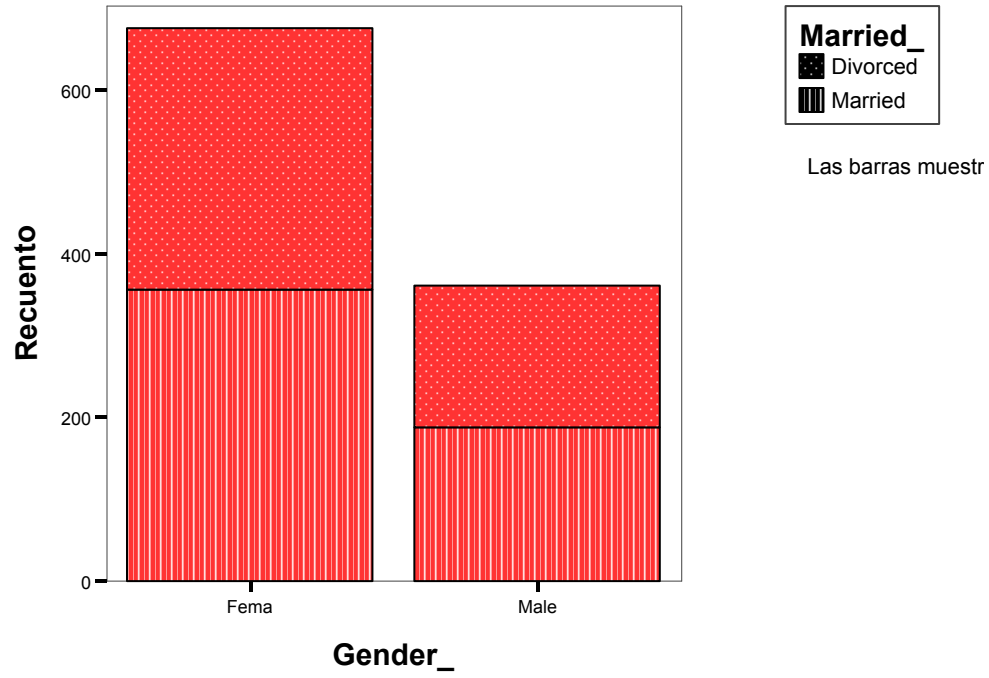
EJERCICIO 2.10.3 ¿El gráfico del Ejercicio 2.10.2 es equivalente a la siguiente tabla?

% de Married_

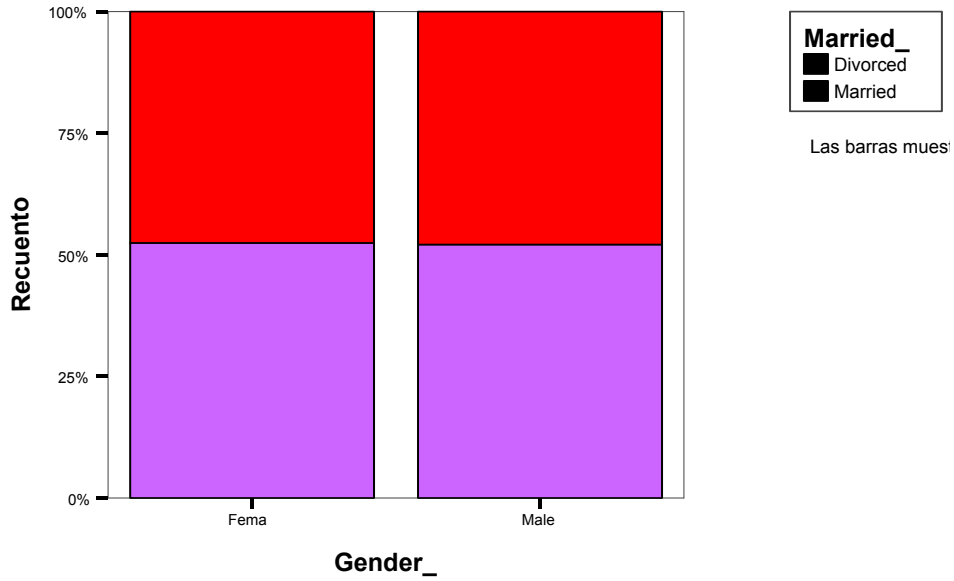
		Married_		
		Divorced	Married	Total
Extramarital_	N	80.2%	95.8%	88.3%
	Y	19.8%	4.2%	11.7%
Total		100.0%	100.0%	100.0%

2.11.Diagramas de barras partidas

- Esta forma de gráfico es también popular para este tipo de datos

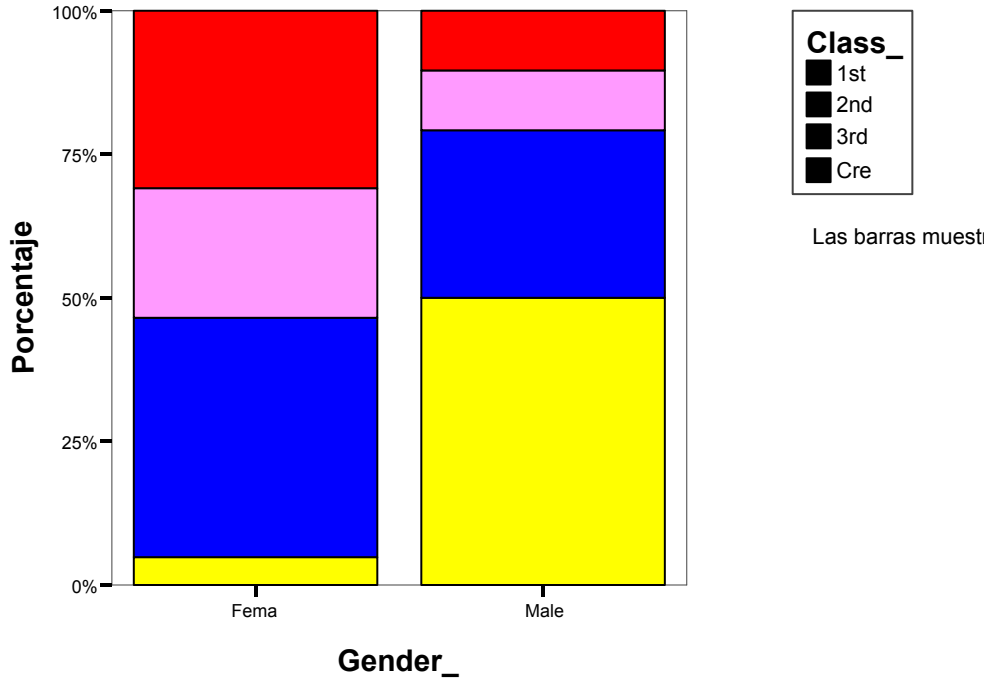


- No obstante, es mejor hacerlo con porcentajes y escalar al 100%



  - Este gráfico se interpreta del siguiente modo. Si los cortes a lo largo del eje horizontal están a la misma altura, entonces no hay diferencias porcentajes dadas las categorías de las barras

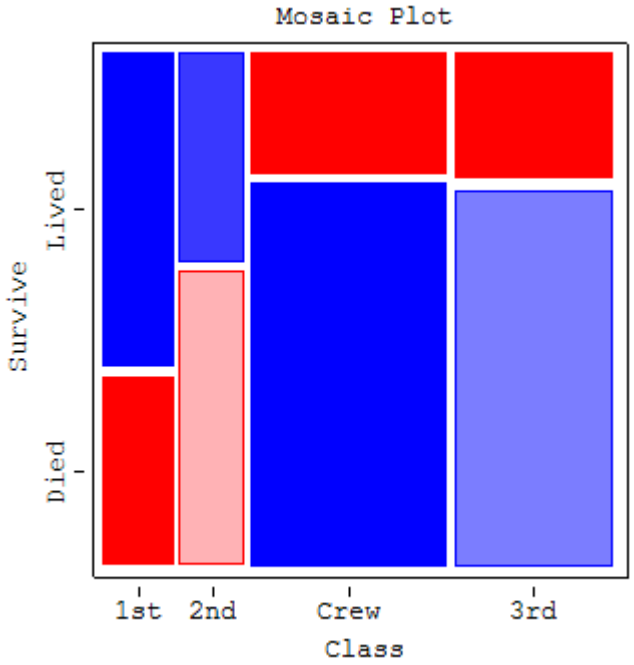
- Este gráfico tiene el inconveniente de que cuando hay diferencias resulta a veces difícil hacer las comparaciones entre categorías ya que no están alineadas



  - (Fijaros en el % de tripulación en los varones ¿Cómo puede influir en conclusiones anteriores?)

2.12.Diagramas de mosaico

- Un gráfico al que se le ha dado mucha importancia en fechas recientes es el siguiente:



– En este gráfico, tanto las columnas como las filas representan porcentajes de la tabla de datos

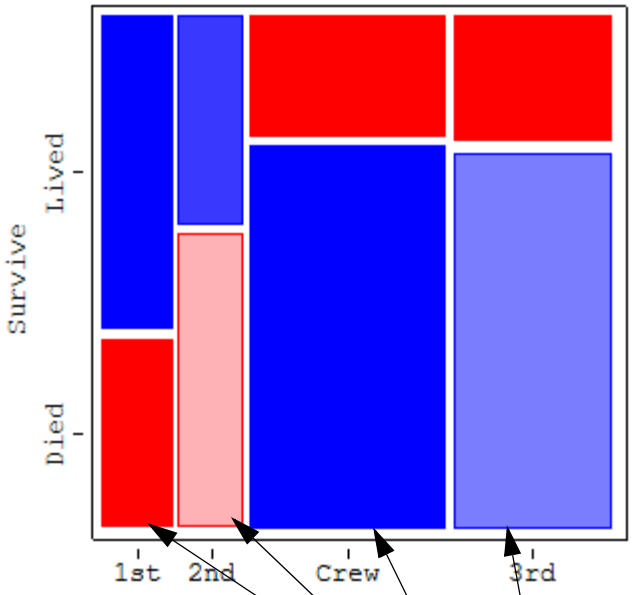


Tabla de contingencia Class_ * Survive_

% de Survive_

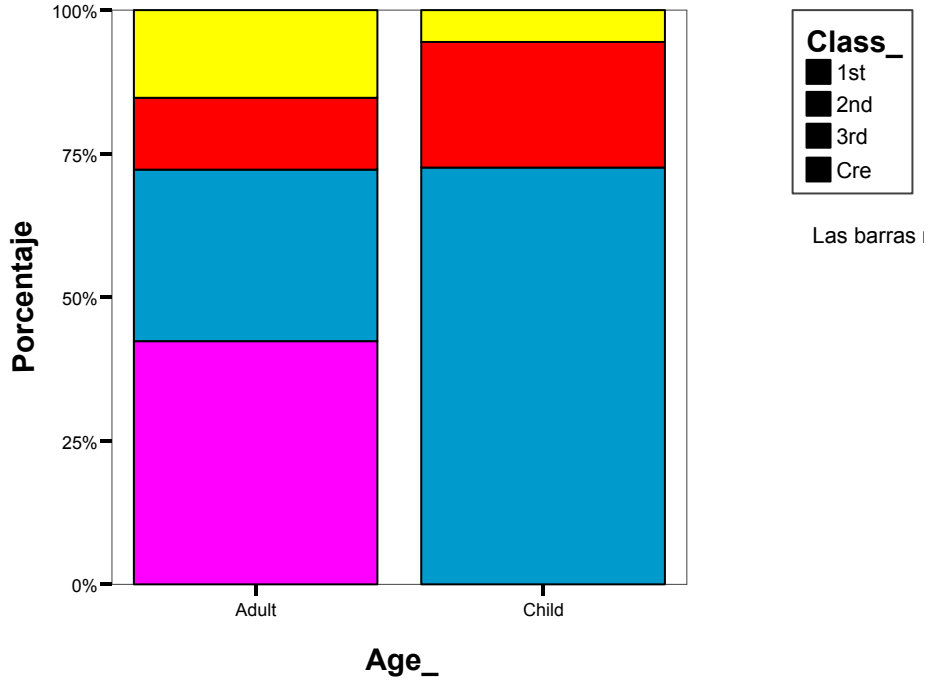
Estos porcentajes de aquí pasan al tamaño de las columnas

				Total
C			5%	14.8%
			5%	12.9%
			0%	32.1%
	Cre	45.2%	29.8%	40.2%
Total		100.0%	100.0%	100.0%

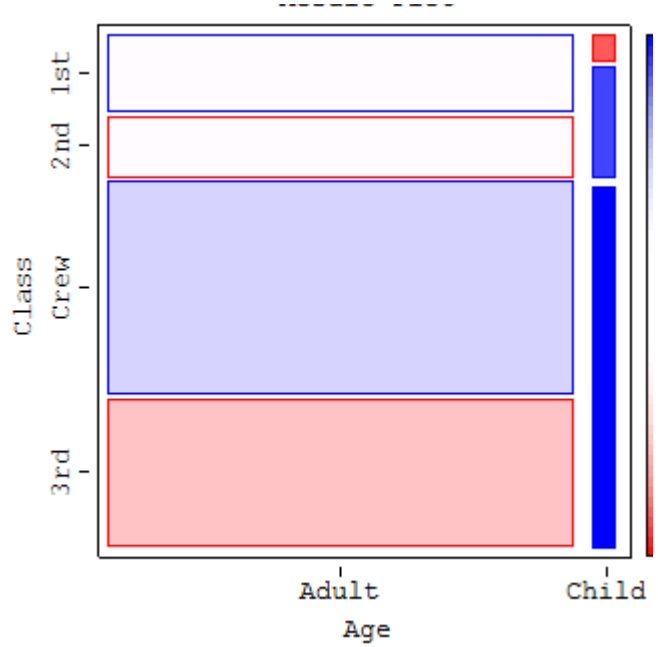
– Esto permite que el tamaño de la celda sea proporcional al tamaño total. Esa información se perdía en el diagrama de barras partidas.

ACTIVIDADES

EJERCICIO 2.12.1 Describe si la edad de los pasajeros tenía algo que ver con el tipo de pasajero



EJERCICIO 2.12.2 Qué aporta este gráfico en comparación con el anterior?



EJERCICIO 2.12.3 El gráfico de mosaico mejora el diagrama de barras aportando el tamaño relativo de una variable. ¿Se podría hacer lo mismo usando gráficos de pastel?

EJERCICIO 2.12.4 En el informe PISA, p. 7 hay un gráfico. Indica los parecidos y diferencias con respecto a los gráficos que hemos estudiado en las últimas secciones ¿Crees que las modificaciones introducidas aportan elementos interesantes al gráfico?

EJERCICIO 2.12.5 En la página 11 del informe PISA, en la parte de abajo, hay un gráfico que ilustra un punto del texto. ¿Qué te parece el uso de ese gráfico? ¿Es razonable o recomendarías hacer otra cosa?

2.13.Más de dos variables y la paradoja de Simpson

"The only statistics you can trust are those you falsified yourself"
W. Churchill

- Hasta ahora hemos visto técnicas centradas en una o dos variables categóricas
- Lo recomendable en general es no ir más allá para evitar complicar demasiado el análisis y/o la presentación de los resultados (a menudo es complicado explicar resultados que afectan a más de dos variables)
- Sin embargo, quedarse en dos variables tiene el peligro de que en nuestros datos se produzca lo que se denomina la paradoja de Simpson.

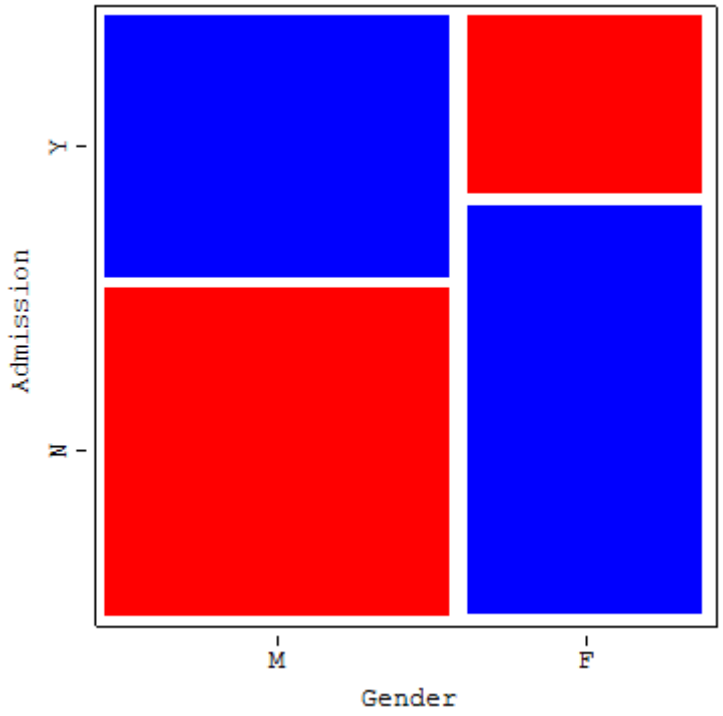
- Un ejemplo muy famoso de esta paradoja ocurrió con el porcentaje de admisiones en la universidad de Berkeley en los años 70. Alguien publicó que en las facultades (allí se entrevista individualmente a los candidatos) se aceptaba al 45% de los hombres y sólo al 30% de las mujeres.

Tabla de contingencia Gender_ * Admission_

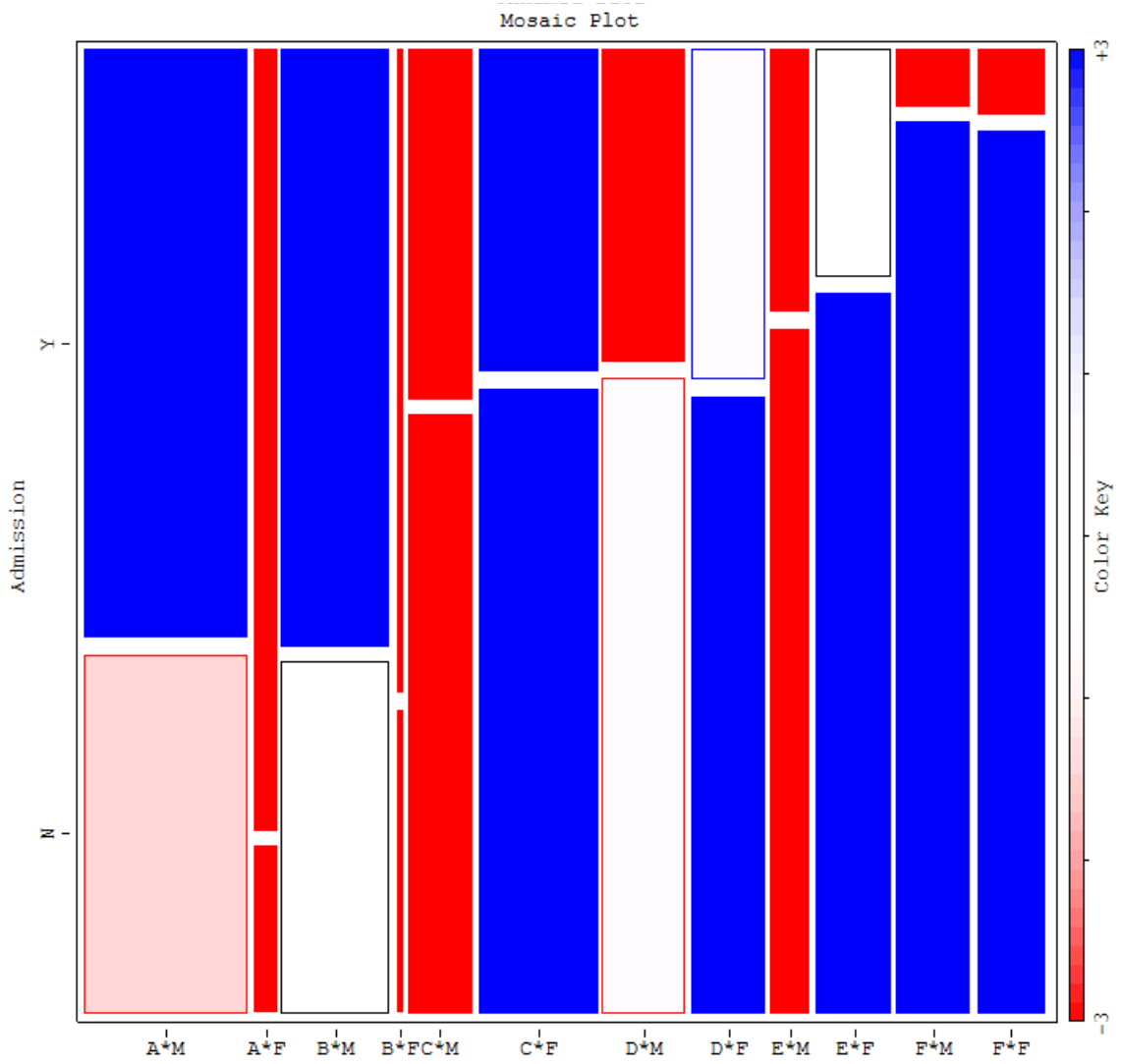
% de Gender_

		Admission_		
		N	Y	Total
Gender_	F	69.6%	30.4%	100.0%
	M	55.5%	44.5%	100.0%
Total		61.2%	38.8%	100.0%

- Viendolo como un gráfico de mosaico tenemos



- Un mosaico muestra que este resultado es engañoso



- Viendolo como una tabla de datos tenemos

Tabla de contingencia Department_ * Admission_ * Gender_

% de Department_

Gender_			Admission_ N	Admission_ Y	Total
F	Department_	A	17.6%	82.4%	100.0%
		B	32.0%	68.0%	100.0%
		C	65.9%	34.1%	100.0%
		D	65.1%	34.9%	100.0%
		E	76.1%	23.9%	100.0%
		F	93.0%	7.0%	100.0%
	Total		69.6%	30.4%	100.0%
M	Department_	A	37.9%	62.1%	100.0%
		B	37.0%	63.0%	100.0%
		C	63.1%	36.9%	100.0%
		D	66.9%	33.1%	100.0%
		E	72.3%	27.7%	100.0%
		F	94.1%	5.9%	100.0%
	Total		55.5%	44.5%	100.0%

Las comparaciones se hacen entre los valores señalados (el primero de arriba con el primero de abajo, el segundo con el segundo, etc.). Se ve que en general los resultados son muy parecidos salvo en el departamento (facultad) A en el que las mujeres son admitidas un 20% más. ¡En realidad la discriminación es a favor de las mujeres no en contra!

- ¿De dónde viene la paradoja?
  - Resulta que las mujeres no solicitaron en los departamentos más fáciles de ser admitidos

Tabla de contingencia Gender_ * Department_

% de Gender_

		Department_						
		A	B	C	D	E	F	Total
Gender_	F	5.9%	1.4%	32.3%	20.4%	21.4%	18.6%	100.0%
	M	30.7%	20.8%	12.1%	15.5%	7.1%	13.9%	100.0%
Total		20.6%	12.9%	20.3%	17.5%	12.9%	15.8%	100.0%

Tabla de contingencia Department_ * Admission_

% de Department_

		Admission_		
		N	Y	Total
Department_	A	35.6%	64.4%	100.0%
	B	36.8%	63.2%	100.0%
	C	64.9%	35.1%	100.0%
	D	66.0%	34.0%	100.0%
	E	74.8%	25.2%	100.0%
	F	93.6%	6.4%	100.0%
Total		61.2%	38.8%	100.0%

2.14.Conclusiones acerca de representación de datos categóricos

- Los datos categóricos son muy comunes y pueden surgir en practicamente cualquier situación práctica o teórica
- En general, el mayor problema con ellos es transmitir los resultados de la manera más clara a otros que los vayan a ver
  - El uso de gráficos y porcentajes son una buena manera de comunicar este tipo de resultados
  - No obstante, antes de comunicar un resultado muy llamativo, comprueba si no son resultado de la paradoja de Simpson

2.15.Valorando las diferencias entre frecuencias (1 variable)

Ejemplo

En este estudio (De Veaux et. al, 2005) se recogió el signo del zodiaco de 256 directivos de empresas que están entre las más grandes del mundo.

Tabla 1: Signo del zodiaco de directivos de empresas entre las más grandes del mundo

Signo	Frecuencia	Porcentaje
Aries	23	9
Tauro	20	8
Geminis	18	7
Cancer	23	9
Leo	20	8
Virgo	19	7
Libra	18	7
Scorpio	21	8
Sagitario	19	7
Capricornio	22	9
Acuario	24	8
Piscis	29	11

2.16.Analizando una variable

El zodiaco y el éxito

- En el ejemplo del zodiaco, lo interesante es ver si hay algun signo que destaca. ¿Cómo podemos ver eso?
  - Empezamos calculando cual es la frecuencia media. Esto es igual a 256/12=21.333.
  - Aquellos signos que tengan más de 21.333 directivos es que estarían relacionados con el éxito, mientras que los que tienen menos no tendrían tanto éxito. Por ejemplo, Acuario o Piscis estarían por encima y Géminis y Libra por debajo.

- No obstante, estos resultados pueden ser debidos a la casualidad (al azar). Una forma de valorar esto es calcular las diferencias entre la frecuencia media (llamada esperada) y la observada y luego sumar:

Tabla 2: Signo del zodiaco de directivos de empresas entre las más grandes del mundo

Signo	Frecuencia	Residual
Aries	23	1.7
Tauro	20	-1.3
Geminis	18	-3.3
Cancer	23	1.7
Leo	20	-1.3
Virgo	19	-2.3
Libra	18	-3.3
Scorpio	21	-0.3
Sagitario	19	-2.3
Capricornio	22	0.6
Acuario	24	2.7
Piscis	29	7.7

– Como las sumas de diferencias respecto a la media son cero elevamos al cuadrado:

$$Suma = \sum (Observada - Esperada)^2$$

– La suma anterior tiene el problema de que su límite es muy amplio. Por ello, se utiliza la siguiente variante:

$$\chi^2 = \sum \frac{(Observada - Esperada)^2}{Esperada}$$

– Si utilizamos la formula anterior, entonces los valores que obtenemos son los siguientes:

Tabla 3: Signo del zodiaco de directivos de empresas entre las más grandes del mundo

Signo	Frecuencia	Residual	Residual²	Residual²/ Esp
Aries	23	1.7	2.8	0.13
Tauro	20	-1.3	1.8	0.08
Geminis	18	-3.3	11.1	0.52
Cancer	23	1.7	2.8	0.13
Leo	20	-1.3	1.8	0.08
Virgo	19	-2.3	5.4	0.25
Libra	18	-3.3	11.1	0.52
Scorpio	21	-0.3	0.1	0.005
Sagitario	19	-2.3	5.4	0.25
Capricornio	22	0.6	0.4	0.02
Acuario	24	2.7	7.1	0.33
Piscis	29	7.7	58.8	2.75

– La suma de la última columna es $\chi^2 = 5,904$

- ¿Cuándo es esa suma grande? La respuesta la tendreis que oir en la segunda parte del curso (pero os anticipo que en este caso los resultados indican que tener ningún signo del zodiaco se da más entre la gente de éxito estudiada).

2.17.Valorando las diferencias entre frecuencias (2 variables)

Ejemplo

En el hundimiento del Titanic se registró los supervivientes y los fallecidos en función de la clase en la que viajaban.

Tabla 4: Fallecidos en el Titanic

	Vivos	Muertos	Suma
Trip.	212	673	885
Primera	202	123	325
Segunda	118	167	285
Tercera	178	528	706
Suma	710	1491	2201

- ¿Qué podemos decir en este caso?
  - ¿Fallecieron más los que estaban en la tripulación?
  - ¿Era lo peor viajar en tercera?
  - ¿Los que estaban en primera lo pasaron mejor?

2.18.Porcentajes para dos variables

- Para analizar este tipo de celdas, en la primera parte del curso calculábamos porcentajes. Por ejemplo, en este caso, si tuvieramos interés en ver si la clase tuvo efecto sobre la supervivencia de los que estaban en el Titanic podríamos calcular lo siguiente:

Tabla 5: Fallecidos en el Titanic porcentajes por fila

	Vivos	Muertos
Trip.	24	76
Primera	62.2	37.8
Segunda	41.4	58.6
Tercera	25.2	74.8

  - En esta tabla podemos ver que efectivamente parece que hay ciertas categorías que fueron más mortales que otras.

- A estos resultados, no obstante, les falta el equivalente de las pruebas de hipótesis que hemos estado calculando en los temas anteriores.
  - Esta prueba de hipótesis nos permitiría decir hasta qué punto lo que observamos en la tabla son significativas desde un punto de vista estadístico.

2.19.Pruebas de hipótesis para tablas de contingencia

- Si en los datos del Titanic se hubiera obtenido el siguiente resultados, diríamos que no hubo efecto en la categoría de tripulante sobre la supervivencia:

Tabla 6: Valores esperados de los fallecidos en el Titanic porcentajes por fila

	Vivos	Muertos	Suma
Trip.	285.5	599.5	885
Primera	104.8	220.2	325
Segunda	91.9	193.1	285
Tercera	227.7	478.3	706
Suma	710	1491	2201

- Si comprobais los totales por fila y por columna vereis que son los mismos que en la tabla anterior, pero los valores de las celdas han variado de modo que son proporcionales a los totales por fila y por columna. Este es el valor esperado y se calcula del siguiente modo:

$$Esperada = \frac{SumaFila \times SumaCol}{Total}$$

  – Por ejemplo, para la casilla Tripulación y Vivos se hace:

$$Esperada = \frac{885 \times 710}{2201} = 285.5$$

- Si ahora calculamos el porcentaje de vivos y muertos para la tabla anterior tenemos lo siguiente:

Tabla 7: Fallecidos en el Titanic porcentajes por fila

	Vivos	Muertos
Trip.	32.3	67.7
Primera	32.3	67.7
Segunda	32.3	67.7
Tercera	32.3	67.7
Total	32.3	67.7

  - Vemos que los porcentajes por categoría de pasajero son iguales a los porcentajes por columna, lo cual significaría que no habría ninguna diferencia en absoluto entre viajar en una clase o en otra en cuanto a la supervivencia

- No obstante, ***en realidad sí que hubo efecto de la clase en la que se viajaba.*** ¿Cómo podemos ver ese efecto?
  - La respuesta está en calcular ***la diferencia entre el valor esperado y el valor observado para cada una de las casillas*** (tal y como hicimos

anteriormente para el ejemplo del zodiaco). Así, con los datos de la ***Table 4 y la Table 6 en un primer paso haríamos***

Tabla 8: Puntuaciones observadas menos esperadas para los datos del Titanic

	Vivos	Muertos
Tripulación	212-285.5 $\overline{=}$ -73.5	673-599.5 $\overline{=}$ 73.5
Primera	202-104.8 $\overline{=}$ 97.2	123-220.2 $\overline{=}$ -97.2
Segunda	118-91.9 $\overline{=}$ 26.1	167-193.1 $\overline{=}$ -26.1
Tercera	178-227.7 $\overline{=}$ -49.7	528-478.3 $\overline{=}$ 49.7

– Fijaros en los valores negativos y positivos. Positivo significa que hubo más de lo esperado (p.e. Vivos en primera y tercera), negativo que hubo menos (p.e. vivos en tripulación y tercera)

- Ahora bien, los valores de la tabla anterior no pueden ser interpretados bien si no tenemos idea de lo grande o lo pequeño que pueden llegar a ser. Una forma de ponerlos en una escala más fácil de entender es dividirlos por la raíz cuadrada del valor esperado ***dando lugar a los residuales tipificados***

Tabla 9: Residuales tipificados

	Vivos	Muertos
Trip.	-4.3	3
Primera	9.5	-6.6
Segunda	2.7	-1.9
Tercera	-3.3	2.3

  - Un residual tipificado es grande si supera un valor de 2 o 2.5. Un 9 como aparece en esta tabla es un valor muy muy muy grande.
  -

- En nuestro caso, hay muchos valores altos. Evidentemente, hubo mucho efecto de la clase sobre la supervivencia
- Por ejemplo, estar en primera llama la atención en que supuso una gran ventaja en cuanto a la supervivencia, mientras que en segunda el efecto no fue tan grande. Tripulación y tercera fueron los más dañados

- Una forma de valorar globalmente el resultado anterior es utilizando el mismo estadístico que en la 2.15.

$$\chi^2 = \sum \frac{(Observada - Esperada)^2}{Esperada}$$

  - Este valor se puede calcular elevando al cuadrado cada una de las casillas de lo que aparece en la Tabla 9 y sumando

Tabla 10: Residuales tipificados

	Vivos	Muertos
Trip.	18.5	9
Primera	90.3	43.6
Segunda	7.3	3.6
Tercera	10.9	5.3

– El resultado es $\chi^2 = 188.4$. Este valor no lo podeis interpretar de momento hasta la segunda parte del curso.

2.20.La V de Cramer

- Podemos mejorar la interpretación de las desviaciones de una tabla a partir de la siguiente fórmula:

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1))}}$$

  - En donde K es el número de filas o columnas (lo que sea más grande)
  - n es el número de casos (la suma total)
  - χ^2 es el valor que hemos calculado anteriormente
- La V es un valor que va entre 0 a 1. Cuanto más alto, más relación hay entre las variables.

- En nuestro caso

$$V = \sqrt{\frac{\dot{188}, 4}{2201 \cdot 3}} = 0, 17$$

ACTIVIDADES

EJERCICIO 2.20.1 Analizando la relación entre Género y Supervivencia tenemos la siguiente tabla. ¿Para quién fue peor este accidente, para los hombres o para las mujeres? ¿Por qué?

Recuento

		Gender_		
		Fema	Male	Total
Survive_	Died	126	1364	1490
	Lived	344	367	711
Total		470	1731	2201

EJERCICIO 2.20.2 ¿Qué dirías de la relación entre género y supervivencia a partir de esta tabla? ¿En qué es diferente de la anterior?

Recuento

		Survive_		
		Died	Lived	Total
Gender_	Fema	126	344	470
	Male	1364	367	1731
Total		1490	711	2201

EJERCICIO 2.20.3 ¿Qué puedes decir de la relación entre Género y Supervivencia a partir de esta tabla?

% de Survive_

		Survive_		
		Died	Lived	Total
Gender_	Fema	8.5%	48.4%	21.4%
	Male	91.5%	51.6%	78.6%
Total		100.0%	100.0%	100.0%

EJERCICIO 2.20.4 ¿Qué podrias decir a partir de esta tabla?

% de Survive_

		Survive_		
		Died	Lived	Total
Gender_	Fema	8.5%	48.4%	21.4%
	Male	91.5%	51.6%	78.6%
Total		100.0%	100.0%	100.0%

EJERCICIO 2.20.5 ¿Y a partir de esta?

% de Gender_

		Survive_		
		Died	Lived	Total
Gender_	Fema	26.8%	73.2%	100.0%
	Male	78.8%	21.2%	100.0%
Total		67.7%	32.3%	100.0%

EJERCICIO 2.20.6 ¿Qué podrias decir de esta relación a partir de estos resultados?

Tabla de contingencia Gender_ * Survive_

Residuos tipificados

		Survive_	
		Died	Lived
Gender_	Fema	-10.8	15.6
	Male	5.6	-8.1

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	456.874[b]	1	.000		
Corrección por continuidad[a]	454.500	1	.000		
Razón de verosimilitudes	434.469	1	.000		
Estadístico exacto de Fisher				.000	.000
Asociación lineal por lineal	456.667	1	.000		
N de casos válidos	2201				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 151.83.

EJERCICIO 2.20.7 Una de las preguntas de la encuesta general de 1991 US es el grado de felicidad que percibían en sus vidas. ¿Qué podrías decir de esa felicidad en función del género?

			Nivel de felicidad			
			Muy feliz	Bastante feliz	No demasiado feliz	Total
Sexo del encuestado	Hombre	% de Sexo del encuestado	32.5%	59.1%	8.4%	100.0%
		Residuos tipificados	.7	.4	-2.0	
	Mujer	% de Sexo del encuestado	30.0%	57.2%	12.9%	100.0%
		Residuos tipificados	-.6	-.3	1.7	
Total		% de Sexo del encuestado	31.1%	58.0%	11.0%	100.0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	7.739[a]	2	.021
Razón de verosimilitudes	7.936	2	.019
Asociación lineal por lineal	4.812	1	.028
N de casos válidos	1504		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 69.44.

EJERCICIO 2.20.8 ¿Y de la felicidad en función de la raza?

Tabla de contingencia Raza del encuestado * Nivel de felicidad

			Nivel de felicidad			
			Muy feliz	Bastante feliz	No demasiado feliz	Total
Raza del encuestado	Blanca	% de Raza del encuestado	32.6%	58.1%	9.3%	100.0%
		Residuos corregidos	2.9	.3	-4.6	
	Negra	% de Raza del encuestado	22.9%	57.7%	19.4%	100.0%
		Residuos corregidos	-2.7	-.1	4.1	
	Otra	% de Raza del encuestado	25.5%	55.3%	19.1%	100.0%
		Residuos corregidos	-.8	-.4	1.8	
Total		% de Raza del encuestado	31.1%	58.0%	11.0%	100.0%

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	24.797[a]	4	.000
Razón de verosimilitudes	22.446	4	.000
Asociación lineal por lineal	16.982	1	.000
N de casos válidos	1504		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5.16.

EJERCICIO 2.20.9 ¿Y la felicidad en función de la zona geográfica en la que viven?

			Nivel de felicidad			
			Muy feliz	Bastante feliz	No demasiado feliz	Total
Región de los Estados Unidos	Nor-Este	% de Región de los Estados Unidos	27.5%	61.2%	11.3%	100.0%
		Residuos corregidos	-2.7	2.3	.4	
	Sur-Este	% de Región de los Estados Unidos	36.3%	52.3%	11.4%	100.0%
		Residuos corregidos	2.7	-2.7	.4	
	Oeste	% de Región de los Estados Unidos	31.7%	58.3%	10.0%	100.0%
		Residuos corregidos	.3	.2	-.7	
Total		% de Región de los Estados Unidos	31.1%	58.0%	11.0%	100.0%

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	10.393[a]	4	.034
Razón de verosimilitudes	10.385	4	.034
Asociación lineal por lineal	2.694	1	.101
N de casos válidos	1504		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 45.09.

EJERCICIO 2.20.10 ¿Y en función de tener una vida excitante, rutinaria o aburrida?

			Nivel de felicidad			
			Muy feliz	Bastante feliz	No demasiado feliz	Total
¿Su vida es excitante o aburrida?	Excitante	% de ¿Su vida es excitante o aburrida?	44.9%	50.2%	4.8%	100.0%
		Residuos corregidos	8.9	-4.7	-5.6	
	Rutinaria	% de ¿Su vida es excitante o aburrida?	19.7%	68.0%	12.3%	100.0%
		Residuos corregidos	-7.4	6.2	1.2	
	Aburrida	% de ¿Su vida es excitante o aburrida?	5.0%	30.0%	65.0%	100.0%
		Residuos corregidos	-3.6	-3.7	11.1	
Total		% de ¿Su vida es excitante o aburrida?	30.4%	58.5%	11.1%	100.0%

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	196.023[a]	4	.000
Razón de verosimilitudes	148.923	4	.000
Asociación lineal por lineal	125.487	1	.000
N de casos válidos	971		

a. 1 casillas (11.1%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 4.45.

Parte III
Representando y describiendo datos numéricos

3.1.Introducción

- En esta sección veremos
  - Como representar datos cuantitativos (1 variable)
  - Como describir datos cuantitativos numéricamente (1 variable)
  - Como representar datos cuantitativos (2 variables)
  - Como describir datos cuantitativos numéricamente (2 variables)
  - Como representar más de dos variables cuantitativas (3 o más)

3.2.Representando datos cuantitativos (1 variable)

"I can't prove it; but I can do more- I can see it"
The innocence of Father Brown. G. K. Chesterton

- Los datos cuantitativos son el caso más importante de datos. La mayoría de los métodos existentes primero fueron desarrollados para datos cuantitativos y luego han sido exportados a otros casos.
- En esta sección veremos
  - Los diagramas de puntos
  - Los histogramas
  - Comparaciones entre variables

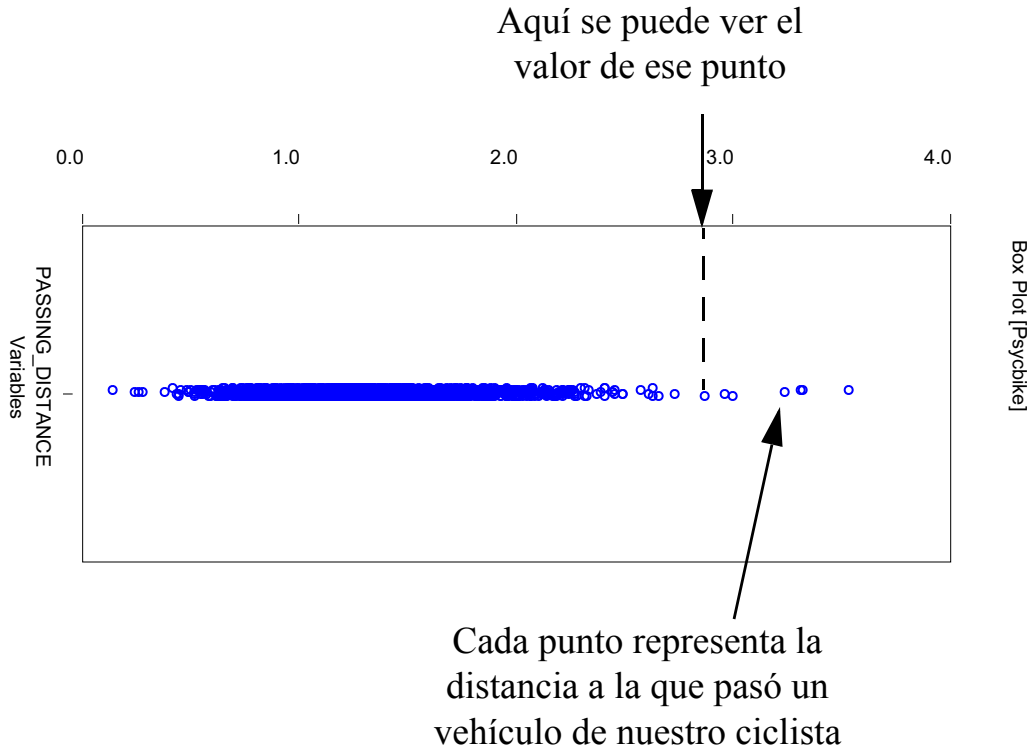
3.3.Diagrama de puntos

Ejemplo

Utilizaremos como un ejemplo la distancia a la que pasan los vehículos cerca de nuestro ciclista

– http://www.drianwalker.com/work.html

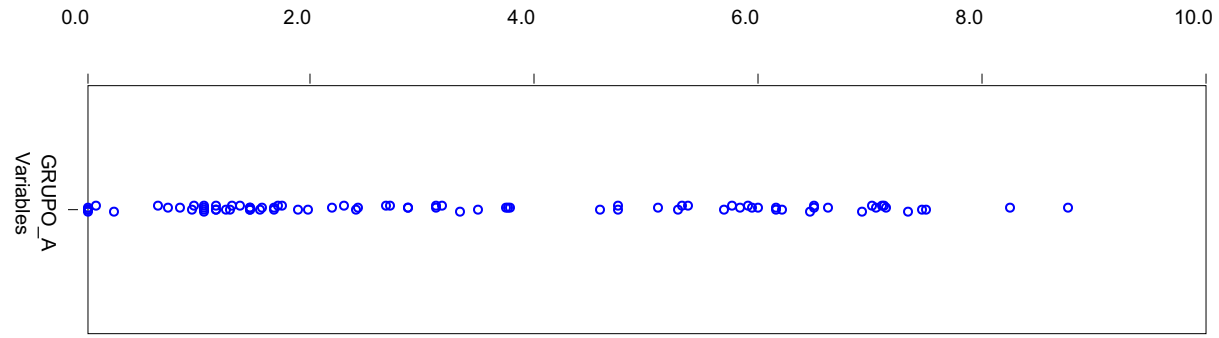
- Un primer gráfico útil para ver esta variable es el siguiente:



- A partir de este gráfico podemos observar
  - cual es la media aproximadamente,
  - los valores más destacados (sobre todo en este caso son interesantes los más cercanos a cero) y
  - si existe algún tipo de huecos, etc.
  - Las distancias entre puntos en algunas partes
- Inconvenientes de este gráfico:
  - Cuando hay muchos datos, los puntos caen todos encima unos de otros y no se aprecia más que un nubarrón (una forma de combatir esto es agitar un poco los puntos pero no siempre es suficiente)

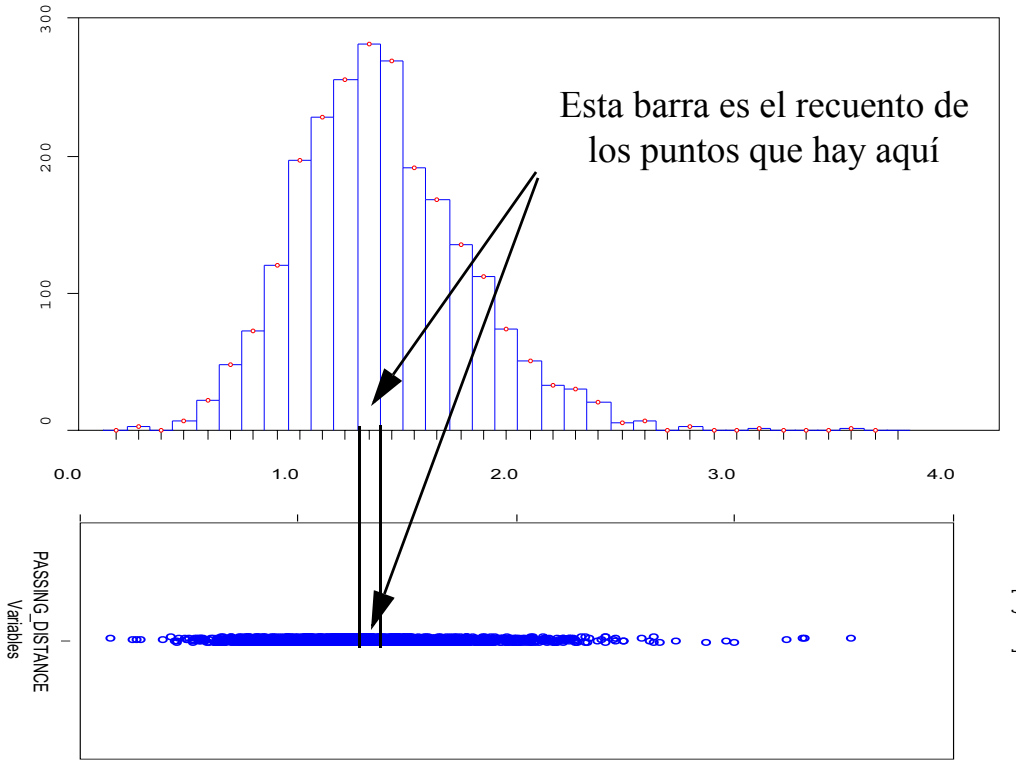
ACTIVIDADES

EJERCICIO 3.3.1 Este es un gráfico de puntos de las notas de un examen de análisis de datos.¿Puedes ver algo interesante en este gráfico?



3.4.Histogramas

- Los histogramas segmentan el diagrama de puntos y cuentan cuantos puntos hay en cada intervalo

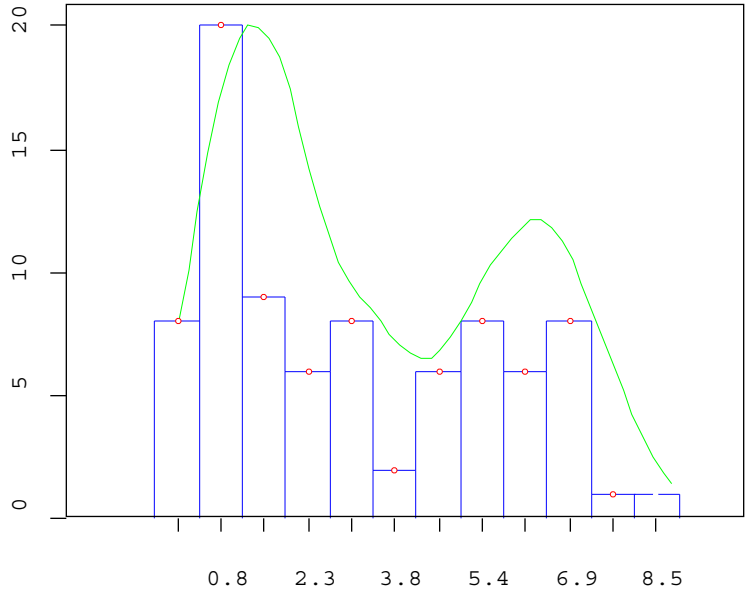


3.5.¿Qué podemos ver en un histograma?

- En los histogramas podemos ver:
  - Si hay una o varias modas
  - Dispersión
  - Si hay simetría o asimetría en los datos
  - Valores destacados (outliers)

3.6.Una o varias modas

- Este es histograma de las notas de análisis de datos de un grupo de hace años.



  - ¿Qué importancia tiene que haya varias modas? Sugiere que hay varios tipos de casos en nuestros datos

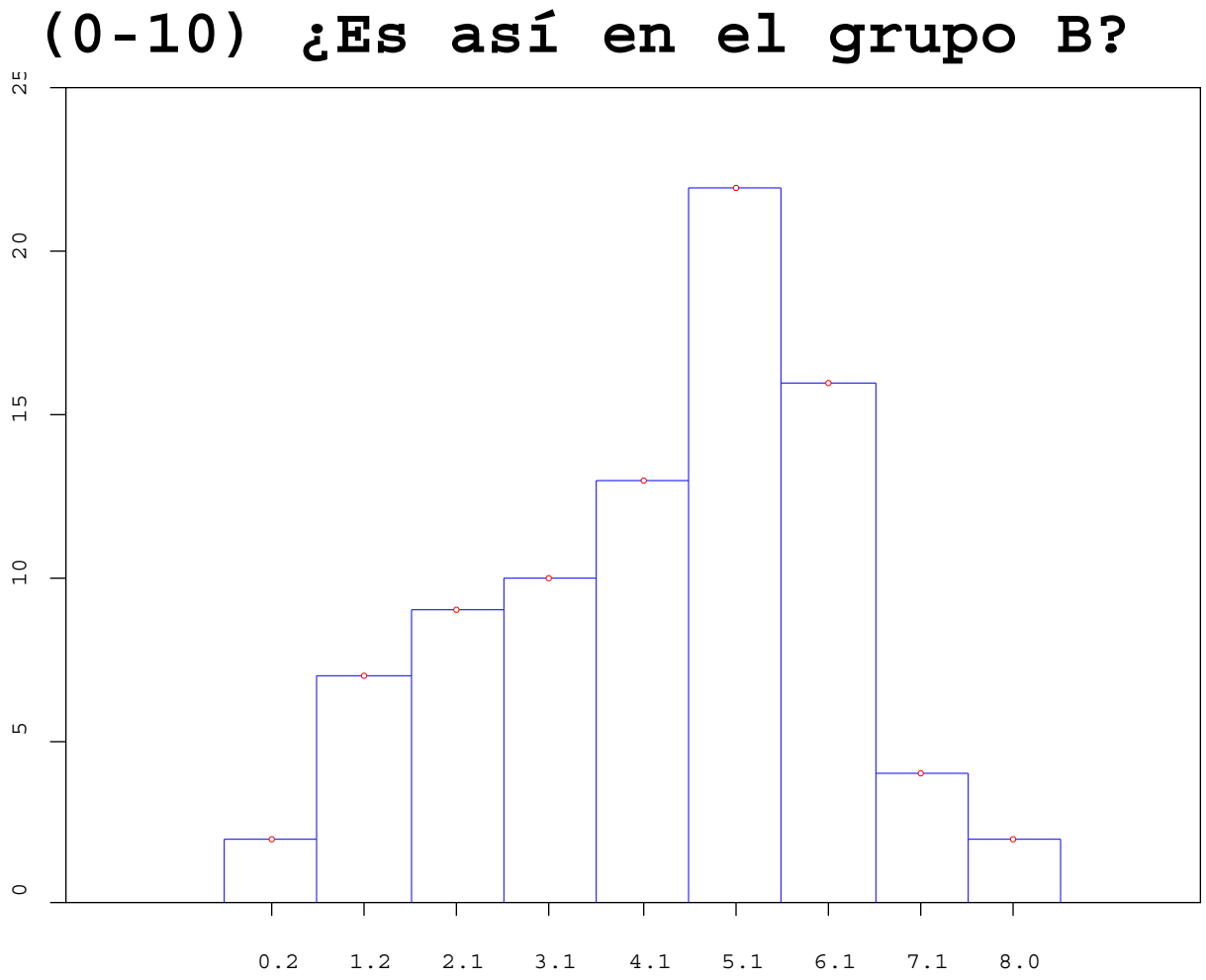
- ¿En este caso qué podríamos concluir? Bueno, es curioso que hay un grupo de sujetos que está centrado en el 6 y otro grupo que está centrado en el 1.5 o en el 2. Para saber qué es lo que les caracteriza necesitaríamos averiguar más cosas pero podemos especular un poco. ¿Alguna idea?
- Otro ejemplo de dos modas: http://Gapminder

3.7.Dispersión

- Por medio de un histograma podemos hacernos una idea de como se produce la dispersión de los datos en una variable
  - La dispersión nos permite valorar los márgenes dentro de los que se mueven los valores de una variable
  - Entender esos márgenes puede ser importante en ciertos casos si tenemos idea de cuales son los límites que deberíamos tener

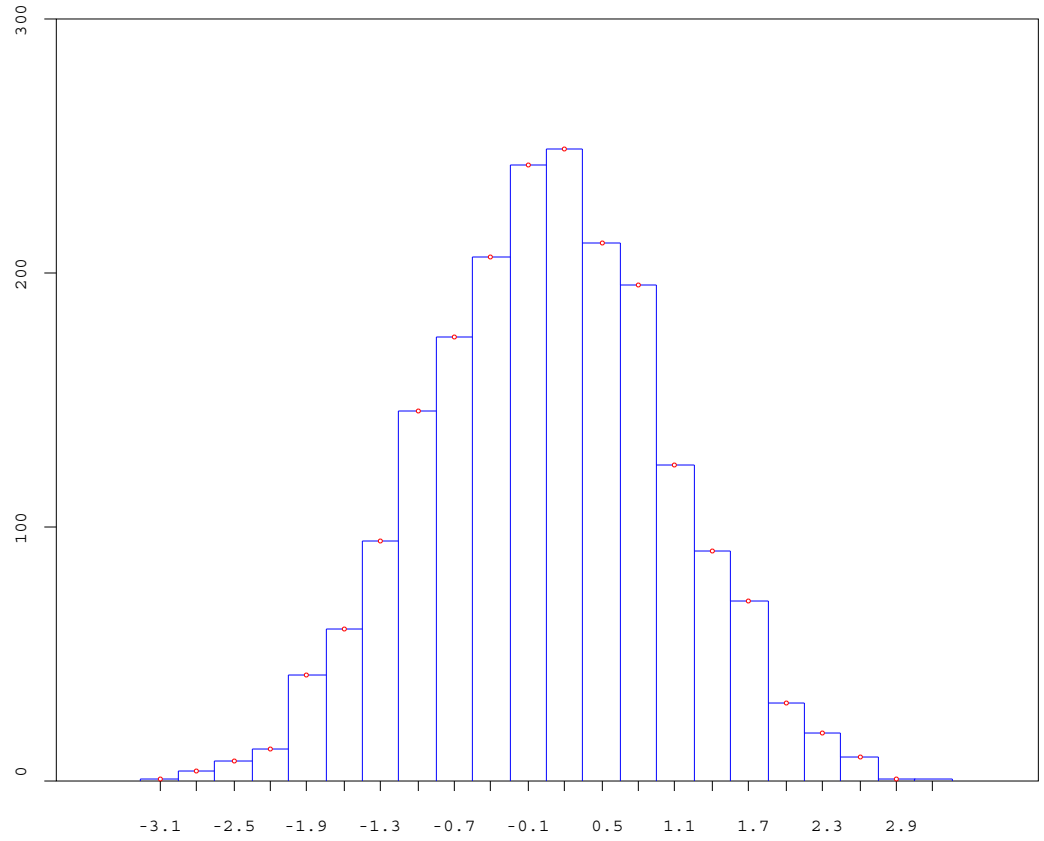
Ejemplo

En los datos de las notas en análisis de datos, parece natural que éstos cubran todos los posibles valores


(0-10) ¿Es así en el grupo B?
0
5
10
15
20
25
0.2
1.2
2.1
3.1
4.1
5.1
6.1
7.1
8.0

3.8.Asimetría/Simetría

– La simetría hace referencia a cuando podemos partir un histograma y doblarlo y ambas partes coincidirían. Este es un ejemplo muuuuy simétrico



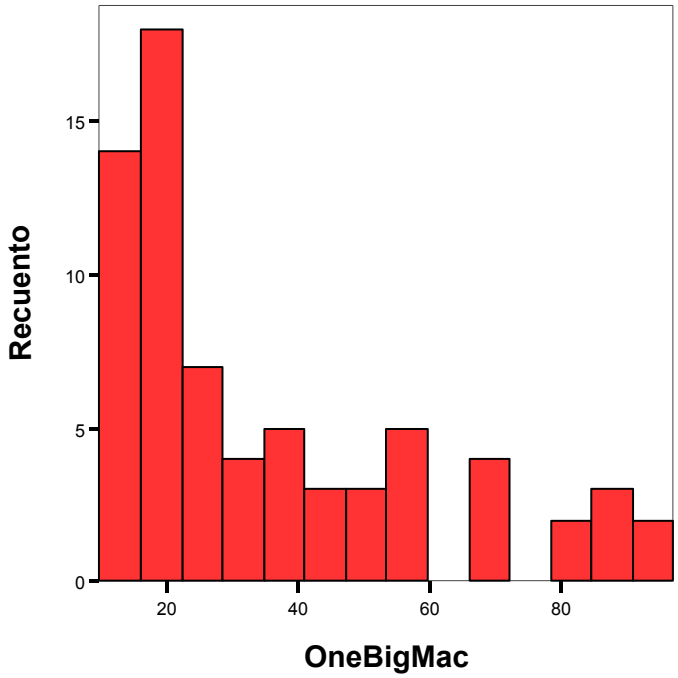
– Sin embargo, cuando trabajamos con datos reales, es habitual que no parezcan tan simétricos.

Ejemplo

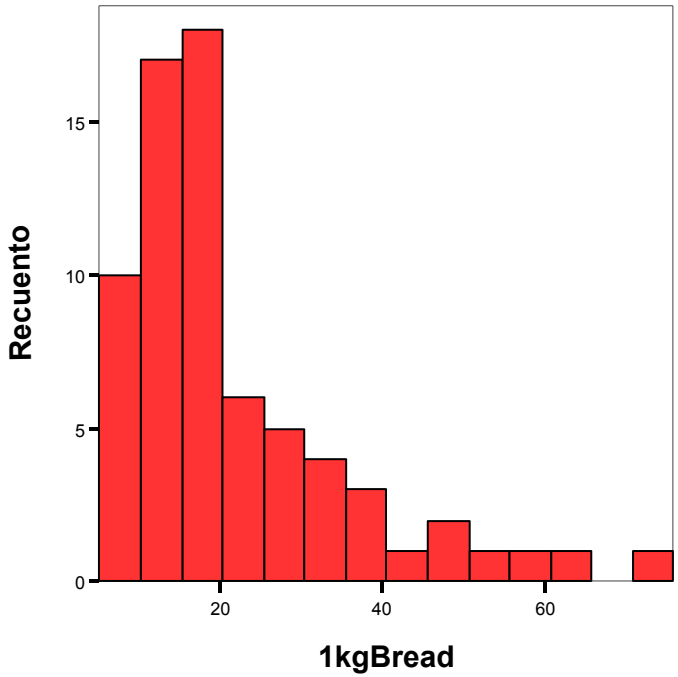
Un banco suizo saca todos los años unas estadísticas que permiten comparar el coste de la vida, el salario cobrado y otros factores a lo largo de las capitales de varios países del mundo. Uno de los índices que es más conocido es el coste en minutos de trabajo (al salario medio) de comprar una hamburguesa Bigmac en todas esas ciudades. A continuación examinaremos esa variable, así como el coste de un kilo de pan y el de un kilo de arroz, siempre en minutos necesarios para adquirir esos productos. Los his-

togramas aparecen a continuación.

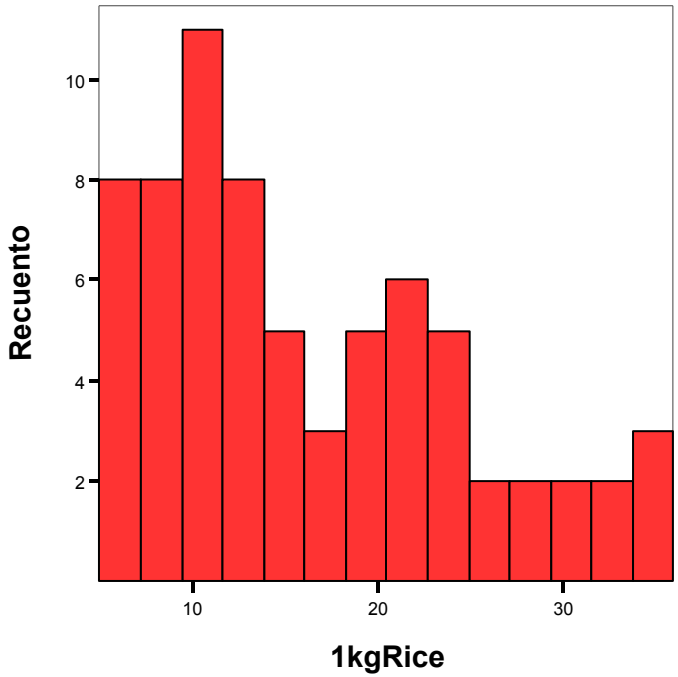
– Histograma para BigMac



– Histograma para Kilo de pan



– Histograma para Kilo de arroz

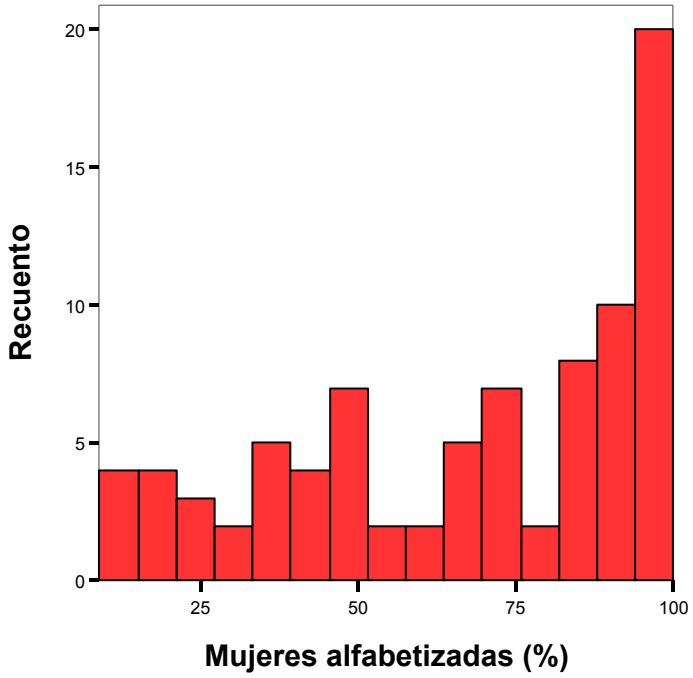


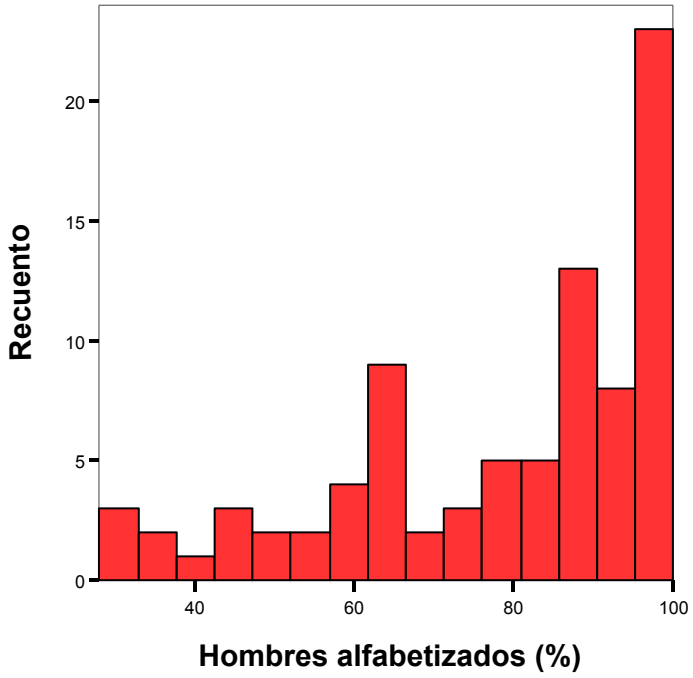
- Como es posible ver en los tres histogramas previos, todos ellos aparecen como asimétricos, con una cola hacia la derecha

  - Este tipo de distribuciones son normales en datos que están limitados por el cero o por un valor mínimo (como el valor mínimo que puede costar una bigmac por ejemplo)
  - Esto se suele dar en datos económicos en los que la mayoría de los casos tienen valores pequeños y a medida que los valores suben desciende la cantidad de casos
- La asimetría contraria es más rara aunque también se puede dar.

Ejemplo

El SPSS proporciona unos datos de ejemplo que denomina Mundo95. Son datos acerca del estado de las naciones del mundo en diversos aspectos socio-económicos (por ejemplo, población, nacimientos, defunciones, alfabetización, etc.). Estos datos son interesantes para explorar la situación de los países del mundo aquel año. Resultados para la alfabetización por género.


Recuento
20
15
10
5
0
25
50
75
100
Mujeres alfabetizadas (%)



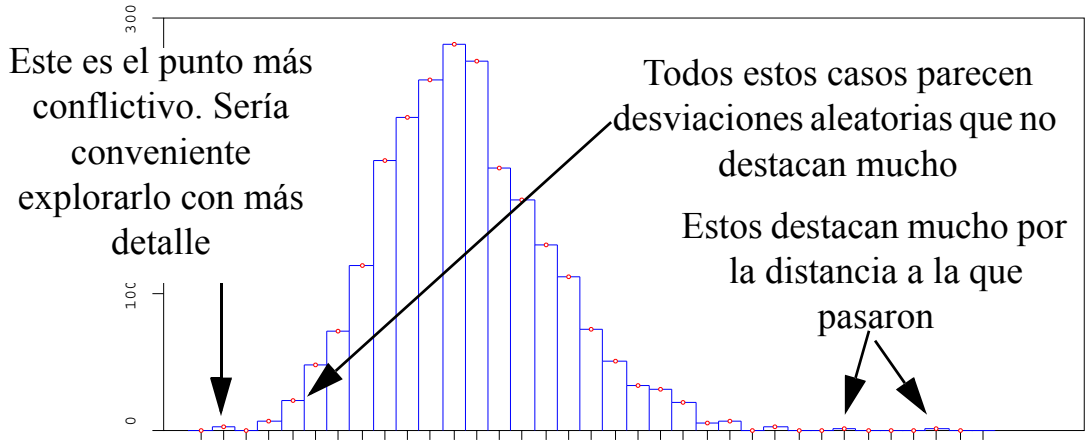
– Observar que el límite está situado en el 100% (no se puede estar más allá de ese valor) y el descenso se produce en la dirección contraria.

3.9.Valores destacados

- ¿Qué es un valor destacado? Uno que destaca, obviamente.
  - En el caso univariado, destacar sólo es posible bien por valores muy altos o por valores muy bajos
  - Cuando tenemos más variables podríamos buscar casos que destacan por la combinación de sus valores (por ejemplo, alguien muy alto y con un peso que sería normal para otros pero que es demasiado bajo para alguien de su estatura).

Ejemplo

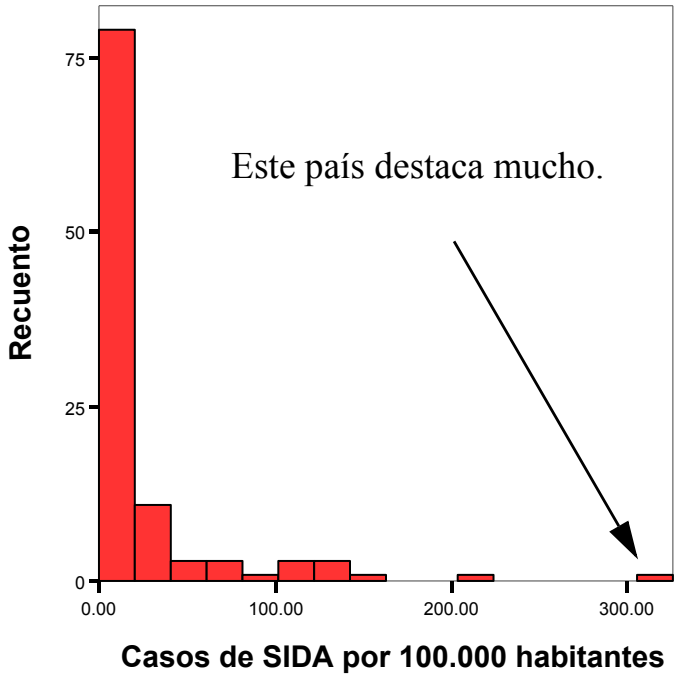
Usaremos los datos de cercanía de coches con el ciclista que explicamos anteriormente. En ese ejemplo, es interesante detectar si existen episodios en que los coches han pasado tan cerca que pueden poner en peligro al ciclista.



Ejemplo

En los datos del mundo podemos encontrar valores más exagerados. En este caso miramos la variable de tasa de sida (casos de sida por cada 100.000

```
hbts)
```


Este país destaca mucho.
Recuento
75
50
25
0
0.00
100.00
200.00
300.00
Casos de SIDA por 100.000 habitantes

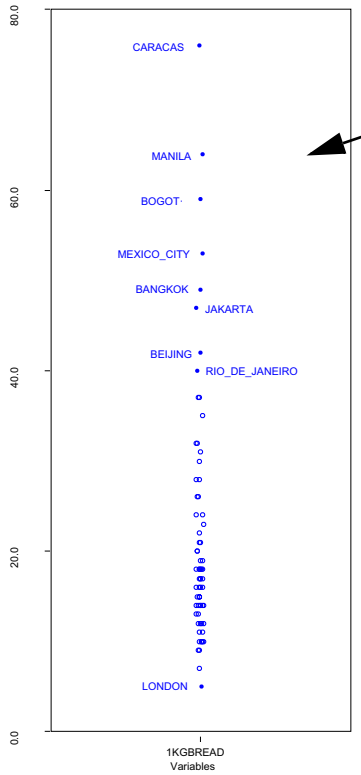
3.10.Identificación: Diagramas de puntos de nuevo

- Identificar datos en gráficos es importante porque permite evaluar casos dentro del contexto de otros casos. Así, se pueden identificar casos con valores y características similares.
- Una de las ventajas más importantes de los diagramas de puntos es que resulta fácil identificar puntos individuales (si se tiene el software adecuado).

Ejemplo

En el ejemplo de bigmac en las ciudades del mundo resulta interesante identificar qué ciudades resulta más costoso adquirir alimentos básicos. Un ejemplo es el siguiente, en el que

se valora el precio del pan



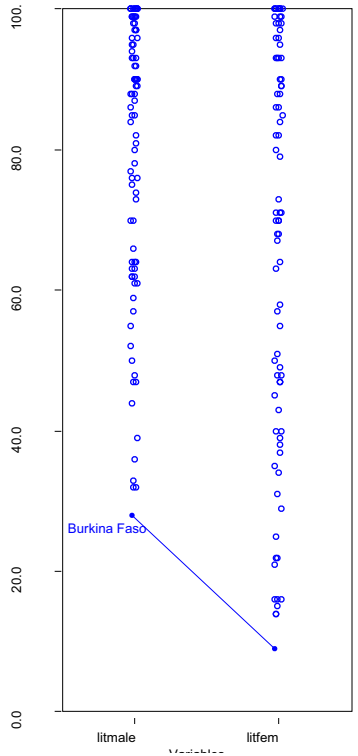
¿Podeis encontrar algo en común a estos países?

3.11.Comparaciones entre variables

- A menudo resulta interesante hacer visualizar varias variables a la vez y ver como cada caso funciona en cada una de las variables.
  - Coches: Velocidad, consumo, precio, etc. son factores que hay que ponderar a la hora de comprarlos
  - Salud: Hay una gran cantidad de parámetros que se pueden considerar para evaluar la salud de alguien. A menudo es conveniente tener varios en cuenta
  - Candidatos: Cuando hay muchos, es importante ver y comparar los diferentes méritos

- Un gráfico que es apropiado para este tipo de situaciones es el de puntos. A continuación tenemos un ejemplo con sólo dos variables: alfabetización femenina y alfabetización masculina. Este gráfico nos permite evaluar los países del mundo en esos aspectos. En el ejemplo de la alfabetización vemos que el mínimo para los hombres es más alto que para las mujeres. Por

ejemplo, la línea que conecta los dos valores de Burkina Faso muestra que en este país los hombres tienen unos porcentajes mucho más altos que las mujeres.



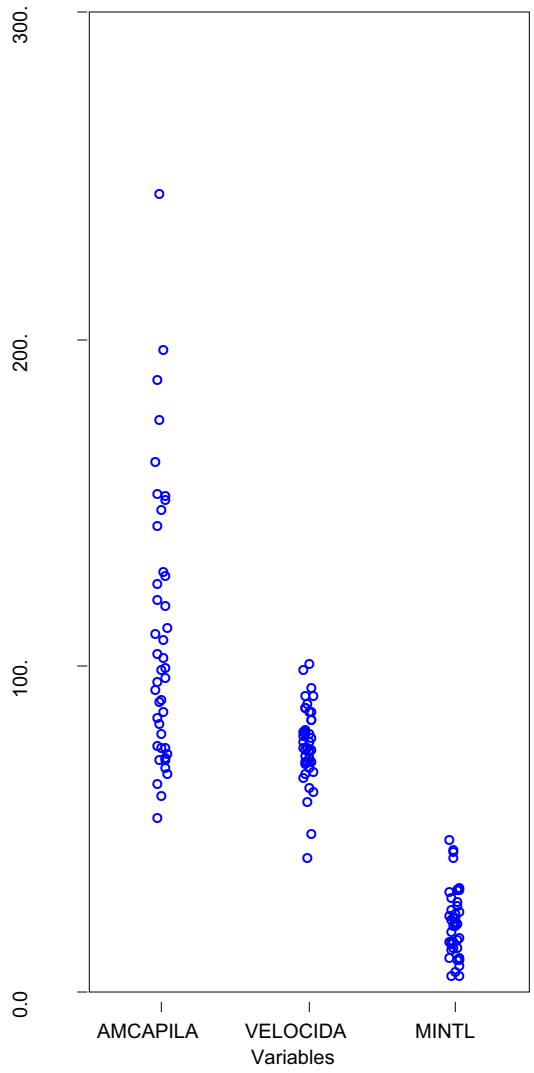
– ¿Creeis que si dibujáramos todas las líneas estas en general estarían horizontales o descenderían?

- Un problema que puede surgir con estas comparaciones es cuando las variables están en escalas diferentes. En ese caso, este gráfico no tiene mucho sentido.

Ejemplo

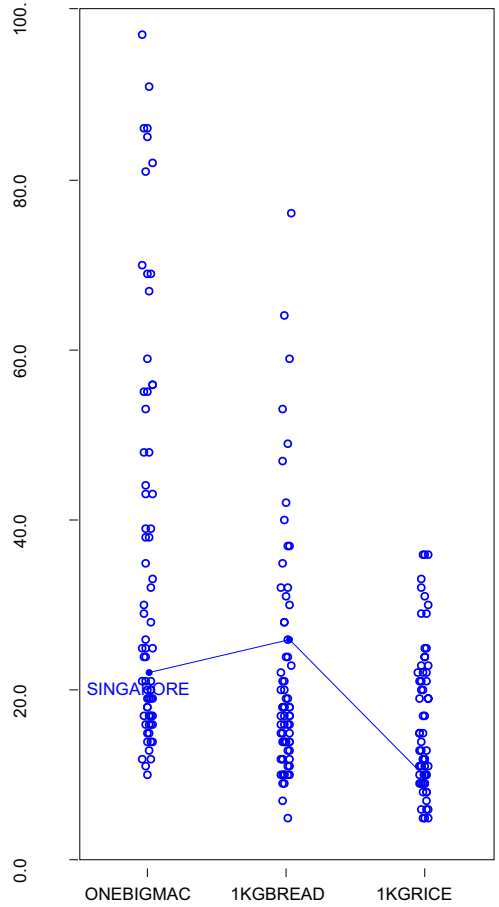
En una investigación realizada por vuestro profesor se analizó el amonio de sujetos enfermos y no enfermos del hígado y su ejecución en la conducción en un simulador (hay sospechas que la gente con problemas de hígado crónicos puede sufrir trastornos cognitivos que les convertirían en peligrosos al volante). Este es un gráfico del amonio (no se en qué medidas), la velocidad a la que condujeron en el simulador (kms/h) y el mintl (una medida de precisión en la conducción que se mide en porcentajes, cuanto más

altos son los valores peor).¿Veis algún problema para la interpretación de éste gráfico? ¿A alguien se le ocurre la manera de hacer este gráfico correctamente?


300.
200.
100.
0.0
AMCAPILA
VELOCIDA
MINTL
Variables

ACTIVIDADES

EJERCICIO 3.11.1 ¿Qué podrías decir de los precios del kilo de arroz, kilo de pan y de una bigmac?



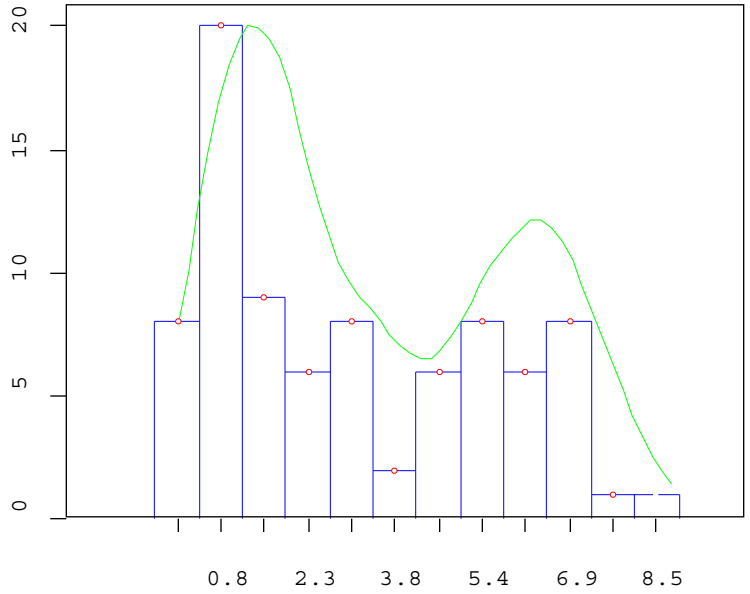
3.12.Descripción de datos numéricos

- Representar gráficamente los datos numéricos está bien, pero por una razón u otra, puede que queramos números para describirlos.
- Los aspectos de interés son:
  - La vulgaridad: Es decir, el centro de los datos
  - La rareza media: Es decir, la dispersión de los datos
  - Las posiciones de los casos individualmente

3.13.La vulgaridad (tendencia central)

"Un estadístico es una persona que puede tener los pies en un horno y la cabeza en un bloque de hielo y decir que, en promedio, se encuentra bien" Chiste anónimo

- ¿Qué diriais que es una nota medianilla en los resultados de la asignatura de Análisis de Datos?



- Posibles respuestas
  - ¿Un cinco? Esa parece una buena nota, no una nota medianilla
  - ¿La media? En este caso la media es 3.46.
  - ¿La nota de la mitad de la gente? (la mediana) En este caso, la mediana es 2.88, que es la nota que tienes a la mitad de la gente por encima o por debajo.
  - ¿La que más gente ha sacado esa nota? (la moda) Eso rondaría el 0.8., o también el 6 si tenemos en cuenta que hay dos modas.

- Supongamos que el profesor está dispuesto a aprobar a todos los que estén por encima de lo medianillo. ¿Qué valor os parecería el correcto?

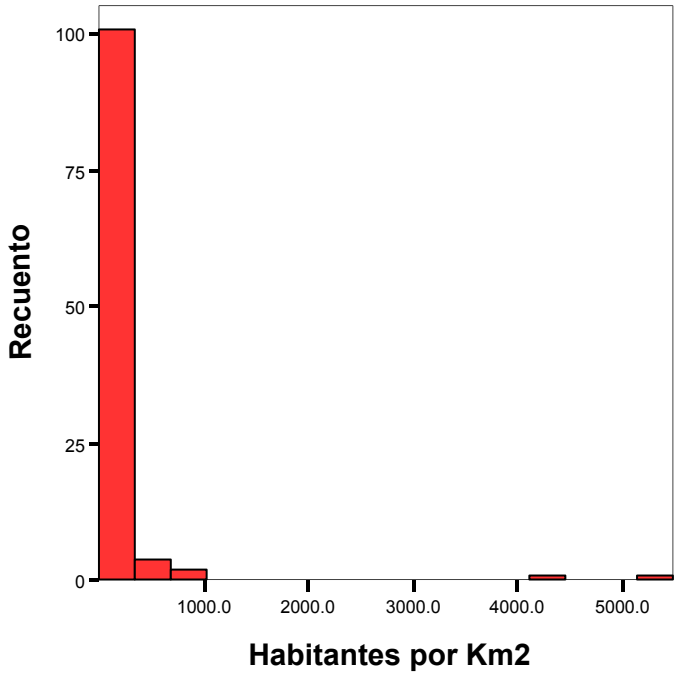
- Cómo hacer el cálculo (por si estais en una isla desierta y os apetece calcular medias y medianas para distraeros)
  - Como calcular la media: (¿de verdad quereis que ponga como calcular la media?)
  - Como calcular la mediana: 1. Ordenar los datos 2. Si el número de casos es impar la mediana es el valor que está la (n+1)/2 posición. Si es par es la media entre el que está en la n/2 y la (n/2)+1 posición.
  - La moda: Cuando los datos son contínuos, es muy difícil que haya repeticiones y hay que agrupar como en el histograma. Si has llegado hasta ahí, es mejor mirarlo en el gráfico.

- ¿Media o mediana?
  - Cuando los datos son simétricos media y mediana coinciden bastante así que no hay conflicto
  - Cuando hay asimetría la media está desplazada en dirección hacia la cola de los datos y la mediana está más centrada. Eso hace que si los datos son muuuuuy asimétricos o hay casos muuuuuy extremos, la media puede resultar en un valor disparatado y la mediana sin embargo resultar razonable. Ese es el caso en que la mediana resulta más útil.

Ejemplo

En los datos acerca del mundo que el SPSS proporciona como ejemplo (Mundo95.sav) tenemos entre otras la variable Densidad de la población (número de habitantes por km2). Un

histograma de esa variable se ve así.



En este caso, la media de habitantes es de 204 habitantes por km2, y la mediana es 63 (los valores extr. son Singapur y Hong Kong)

- ¿Y la moda qué?
  - Si los datos son simétricos, todavía puede ser que los datos sean bimodales
  - Determinar si hay más de una moda puede indicar que hay varios grupos en los datos. Si hay varios grupos, resulta interesante dar las medias y las medianas de los grupos por separado.

3.14.Robustez y medidas de tendencia central alternativas

- Cuando un estimador de tendencia central se ve muy afectado por valores extremos se dice que no es robusto
- Un estimador de tendencia central es robusto cuando unos pocos valores extremos no afectan demasiado a su resultado
- A lo largo de los años se han propuesto métodos de estimación de la tendencia central robustos (uno de ellos era la mediana). Aquí veremos otros dos métodos: medias recortadas y medias ponderadas

- Medias recortadas.
  - La idea de estas medias es excluir un porcentaje de valores a la hora de hacer el cálculo tanto a izquierda como a derecha
  - Se suele hablar del porcentaje recortado a cada lado. Por ejemplo, recortado al 5% o al 10% que significa que se han recortado el 10% o el 20% respectivamente del total de los datos
  - Rosemberg and Gasco (1983) dicen que para datos que tengan histogramas que pueden ir desde simétricos hasta bastante asimétricos se puede aplicar las siguientes reglas para estimar la

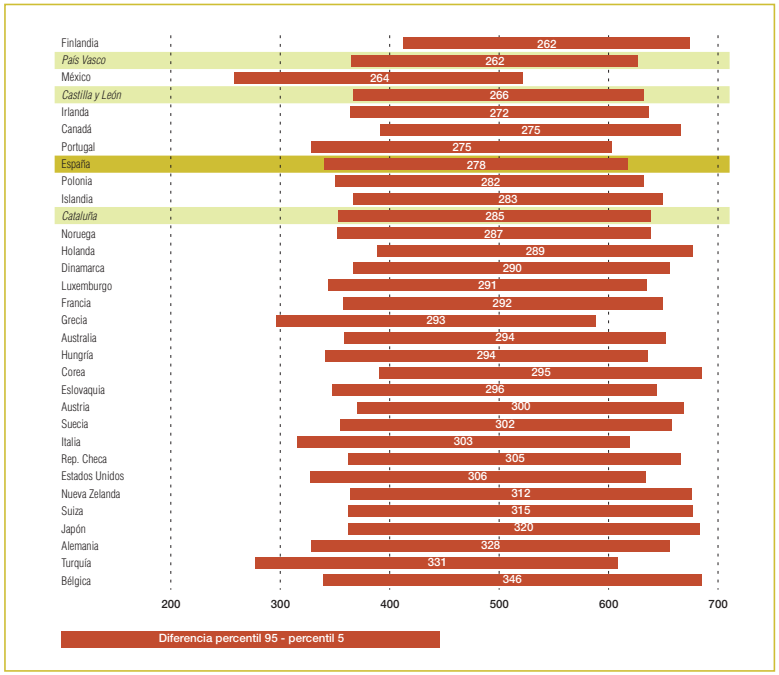
tendencia central: para n<=6 usar la mediana, para n=7 quitar dos observaciones, para n>=8 recortar 25% de cada lado

- Medias ponderadas
  - Las medias ponderadas calculan la media de los datos dando más peso a las puntuaciones centrales y menos a las laterales
  - Hay cuatro formas alternativas en que estas medias se pueden calcular (dejaremos los detalles al ordenador)
  - De ellas, Hubers no debe ser utilizado cuando hay valores muy extremos en la muestra. Los otros tres (Hampels, Andrews y Biweight) son apropiados aunque haya valores extremos pero pueden dar resultados diferentes por lo que habrá que valorarlos en conjunto

3.15.La media de la dispersión

- En el informe PISA se considera como un criterio importante para valorar un sistema educativo no sólo que la media sea alta sino también que tenga pocas desigualdades (definidas como distancia entre el percentil 5 y el 95).

Dispersión de los resultados en Matemáticas



Los países están ordenados de menor a mayor dispersión entre entre los percentiles 5 y 95

- España se encuentra colocada en la parte alta de la clasificación, como un país en donde las diferencias internas de resultados son relativamente menores. Es digna de destacar la posición del País Vasco y de Castilla y León como territorios más equitativos que la media española y que la mayoría de los países de la OCDE.
- Los países con sistemas educativos segregadores y con itinerarios formativos –Bélgica, Alemania, Suiza– producen mayores dispersiones en sus resultados mientras que los países con un sistema educativo más integrador y comprensivo, España entre ellos, tienden a ofrecer menor dispersión.
- Los países que logran aunar excelencia y equidad presentan en el gráfico anterior barras cortas y situadas más a la derecha, en la zona de las puntuaciones más altas. Es el caso, por ejemplo, de Finlandia y Canadá. España presenta una barra corta, pero no queda suficientemente situada en la zona de puntuaciones altas: se encuentra aún falta de excelencia, aunque no de equidad.
- El gráfico siguiente presenta la misma situación en un formato distinto, más puntual. La excelencia sigue estando representada por los promedios de las puntuaciones en Matemáticas y la equidad por las desviaciones típicas de esas mismas puntuaciones.

- En el informe PISA miden la variación utilizando percentiles. No obstante, la forma más común de hacerlo es utilizando la desviación típica

Ecuación (1)

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n}}$$

- Un ejemplo de cálculo, desviación típica entre el que se ha comido un pollo y el que se ha quedado sin comer

Datos

– Tenemos dos sujetos: uno se ha comido un pollo, el otro cero

Sujeto	Pollo
1	1
2	0

– Primero calculamos la media:

$$\bar{y} = \frac{1+0}{2} = 0.5$$

– Luego, hacemos el sumatorio (el símbolo $\sum_{i=1}^{n}$ quiere decir que para cada caso empezando por el primero $i = 1$ hasta el último(n) haz lo que pone en la fórmula y luego haz la suma. Así, para cada valor de la variable hacemos la resta con la media y lo elevamos al cuadrado:

$$Sujeto1 \Rightarrow (1 - 0.5)^2 = 0.25$$

$$Sujeto2 \Rightarrow (0 - 0.5)^2 = 0.25$$

– Luego sumamos

$$0.25 + 0.25 = 0.5$$

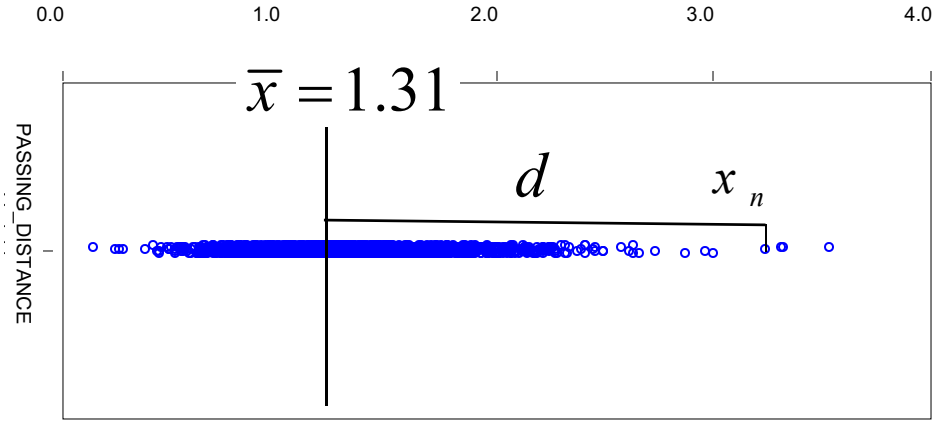
– Luego dividimos por n (en este caso 2)

$$\frac{0.5}{2} = 0.25$$

– Finalmente, sacamos la raíz cuadrada

$$\sqrt{0.25} = 0.5$$

- ¿Qué es lo que hace la fórmula de la desviación típica?



1. Calcular las distancias entre la media y cada punto
2. Elevar cada distancia al cuadrado
3. Sumar
4. Sacar la raíz cuadrada

3.16.Variabilidad y número de datos

- Un hecho bastante común es que el número de casos disponible está relacionado con la variabilidad. A mayor número de casos, más variabilidad.
- Esto en principio resulta un poco inesperado ya que la desviación típica en el fondo es una media (de desviaciones, pero una media al fin y al cabo), y las medias, como dividen por el número de casos no se deberían ver afectadas por el tamaño del conjunto de datos.
  - No obstante, cuantos más casos hay, ***más posibilidades hay para que haya más variabilidad.***

Ejemplo

En un congreso de hace unos años, se organizó un concurso estadísticos. La idea era coger unos datos y hacer el análisis más interesante sobre ellos. En este caso, se cogieron los datos nutricionales de una serie de cereales para el desayuno como los que había en un supermercado concreto. Las variables recogidas son las calorías de este tipo de cereales, las proteínas, grasa y así. En este caso, analizaremos la variabilidad de los cerereales en función de la marca. Ahora bien, hay que tener en cuenta que hay marcas

que comercializan más cereales que otros.

– Resultados para marca N

VARIABLES (Numeric)	MEAN	StDv	VARIANCE	SKEWNESS	KURTOSIS	N
calories	86.67	10.33	106.67	-0.67	0.59	6.0
protein	2.83	0.75	0.57	0.31	-0.10	6.0
fat	0.17	0.41	0.17	2.45	6.00	6.0
sodium	37.50	54.93	3017.50	1.26	0.10	6.0
fiber	4.00	3.10	9.60	1.88	4.23	6.0
carbohydrates	16.00	5.87	34.40	-1.66	2.98	6.0
sugars	1.83	2.86	8.17	1.02	-1.56	6.0
potassium	120.67	91.83	8432.67	0.85	2.27	6.0
vitaminerals	8.33	12.91	166.67	0.97	-1.88	6.0

– Resultados para marca K

VARIABLES (Numeric)	MEAN	StDv	VARIANCE	SKEWNESS	KURTOSIS	N
calories	108.64	22.74	517.10	-0.32	2.03	22.0
protein	2.68	1.09	1.18	1.20	3.13	22.0
fat	0.64	0.85	0.72	1.33	1.46	22.0
sodium	177.05	85.39	7292.05	-0.53	0.08	22.0
fiber	2.82	3.25	10.54	2.38	6.44	22.0
carbohydrates	15.32	4.48	20.04	-0.23	-0.86	22.0
sugars	7.27	4.38	19.16	0.13	-1.21	22.0
potassium	106.36	92.73	8598.05	1.34	1.02	22.0
vitaminerals	35.23	26.34	693.99	2.28	3.50	22.0

– Podemos ver que en general la marca K, que tiene más cereales, tiene más variabilidad que la N en la mayoría (pero no todas) las variables

– No obstante, el resultado anterior no está garantizado que ocurra. Por ejemplo, comparando la marca Q (arriba) con la K de nuevo (abajo)

VARIABLES (Numeric)	MEAN	StDv	VARIANCE	SKEWNESS	KURTOSIS	N
calories	95.00	29.28	857.14	-1.02	-0.52	8.0
protein	2.63	1.60	2.55	0.26	-1.74	8.0
fat	1.75	1.58	2.50	1.12	2.27	8.0
sodium	92.50	99.50	9900.00	0.30	-2.11	8.0
fiber	1.34	1.00	1.00	-0.28	-1.30	8.0
carbohydrates	10.00	4.81	23.14	-2.09	4.73	8.0
sugars	5.25	5.09	25.93	-0.02	-1.70	8.0
potassium	74.38	43.38	1881.70	0.03	-1.72	8.0
vitaminerals	12.50	13.36	178.57	0.00	-2.80	8.0

VARIABLES (Numeric)	MEAN	StDv	VARIANCE	SKEWNESS	KURTOSIS	N
calories	108.64	22.74	517.10	-0.32	2.03	22.0
protein	2.68	1.09	1.18	1.20	3.13	22.0
fat	0.64	0.85	0.72	1.33	1.46	22.0
sodium	177.05	85.39	7292.05	-0.53	0.08	22.0
fiber	2.82	3.25	10.54	2.38	6.44	22.0
carbohydrates	15.32	4.48	20.04	-0.23	-0.86	22.0
sugars	7.27	4.38	19.16	0.13	-1.21	22.0
potassium	106.36	92.73	8598.05	1.34	1.02	22.0
vitaminerals	35.23	26.34	693.99	2.28	3.50	22.0

– La marca Q tiene más variabilidad en calorias, en proteinas, en grasa y en otras cosas que la K

3.17.Desviación típica y datos asimétricos

- Ya vimos que cuando los datos son asimétricos, la media puede dar resultados extraños y por tanto es conveniente usar medianas. Lo mismo puede pasar perfectamente con la desviación típica por lo que a menudo es conveniente usar una medida diferente.
- La medida alternativa es el rango intercuartil.
  - El rango intercuartil es la distancia entre la puntuación que deja por debajo de sí el 25% de las puntuaciones (el primer cuartil) y la que deja el 75% (el tercer cuartil)
  - Entre el primer cuartil y el tercer cuartil están el 50% de las puntuaciones.

- Veamos los rangos intercuartiles con los datos de los cereales

– Para los cereales K

VARIABLES (Ord. & Num.)	MINIMUM	1st Q	MEDIAN	3rd Q	MAXIMUM
calories	50.00	100.00	110.00	115.00	160.00
protein	1.00	2.00	3.00	3.00	6.00
fat	0.00	0.00	0.00	1.00	3.00
sodium	0.00	140.00	180.00	225.00	320.00
fiber	0.00	1.00	1.50	3.00	14.00
carbohydrates	7.00	13.50	15.00	19.00	22.00
sugars	0.00	3.00	7.00	11.50	15.00
potassium	20.00	37.50	75.00	145.00	330.00
vitaminerals	25.00	25.00	25.00	25.00	100.00

VARIABLES (Numeric)	IQ-RANGE	RANGE	MID-RANGE
calories	15.00	110.00	105.00
protein	1.00	5.00	3.50
fat	1.00	3.00	1.50
sodium	85.00	320.00	160.00
fiber	2.00	14.00	7.00
carbohydrates	5.50	15.00	14.50
sugars	8.50	15.00	7.50
potassium	107.50	310.00	175.00
vitaminerals	0.00	75.00	62.50

– Para los cereals N

VARIABLES (Ord. & Num.)	MINIMUM	1st Q	MEDIAN	3rd Q	MAXIMUM
calories	70.00	85.00	90.00	90.00	100.00
protein	2.00	2.50	3.00	3.00	4.00
fat	0.00	0.00	0.00	0.00	1.00
sodium	0.00	0.00	7.50	47.50	130.00
fiber	1.00	3.00	3.00	3.50	10.00
carbohydrates	5.00	15.50	17.50	19.50	21.00
sugars	0.00	0.00	0.00	2.50	6.00
potassium	-1.00	92.50	107.50	130.00	280.00
vitaminerals	0.00	0.00	0.00	12.50	25.00

VARIABLES (Numeric)	IQ-RANGE	RANGE	MID-RANGE
calories	5.00	30.00	85.00
protein	0.50	2.00	3.00
fat	0.00	1.00	0.50
sodium	47.50	130.00	65.00
fiber	0.50	9.00	5.50
carbohydrates	4.00	16.00	13.00
sugars	2.50	6.00	3.00
potassium	37.50	281.00	139.50
vitaminerals	12.50	25.00	12.50

-

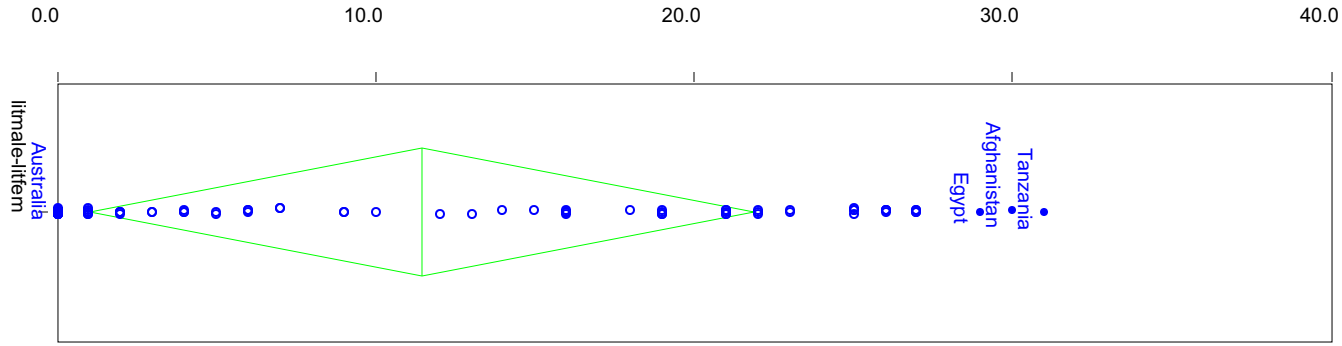
– No obstante, equivalente no significa el mismo valor. Estos valores no coinciden como pasaba con la media y la mediana

3.18.Teniéndolo todo

- Ya hemos visto que los gráficos estadísticos permiten tener una visión detallada de los datos, mientras que los resúmenes numéricos (media, desviación típica) permiten más precisión
  - ¿No sería interesante tenerlo todo?
- Para mezclar ambas perspectivas utilizaremos el diagrama de puntos. Podemos hacer dos versiones:
  - una con medias y desviaciones típicas
  - otra con medianas y rangos intercuartiles (y algunas cosas más que ya veremos)

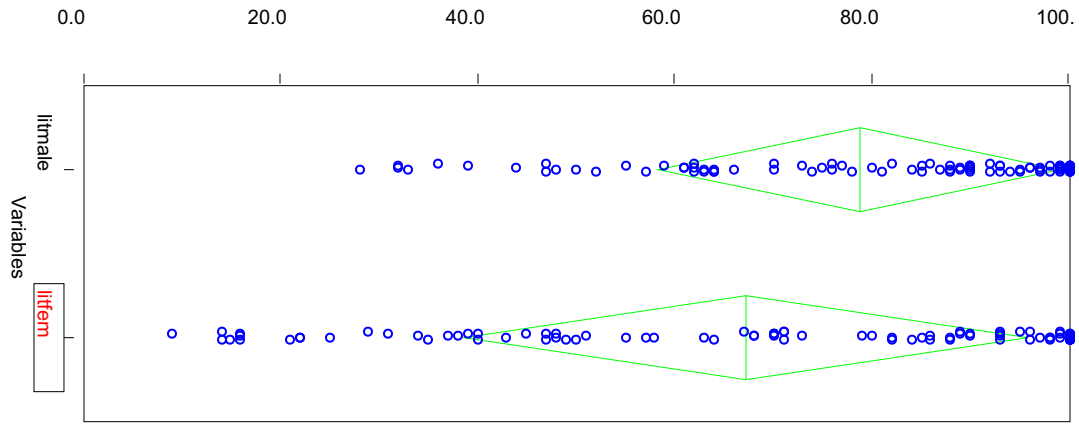
3.19.Gráficos de puntos con medias y desviaciones típicas

- Antes vimos gráficos que comparaban la alfabetización masculina y femenina países del mundo. Otra forma de ver esos datos es restando la alfabetización de un género respecto del otro.



La línea central indica la media (vemos que es aproximadamente 12% y las puntas del diamante nos indican una desviación típica arriba y otra abajo.

- Este tipo de gráficos son interesantes para comparar diferentes grupos o variables. Por ejemplo, si ponemos la alfabetización masculina y femenina juntas tenemos:

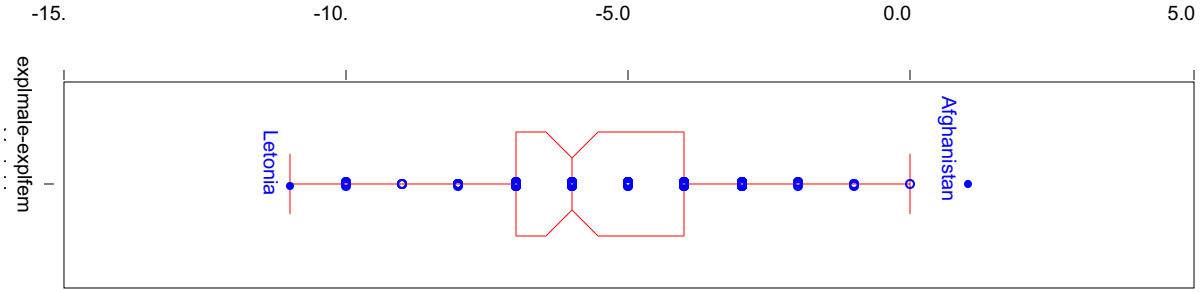


  - Es fácil ver que la media de los hombres es más alta que la de las mujeres así como otras cosas.
  - También, la desviación típica es mayor en las mujeres: ¿alguna explicación a esto?

3.20.Diagramas de cajas (y bigotes)

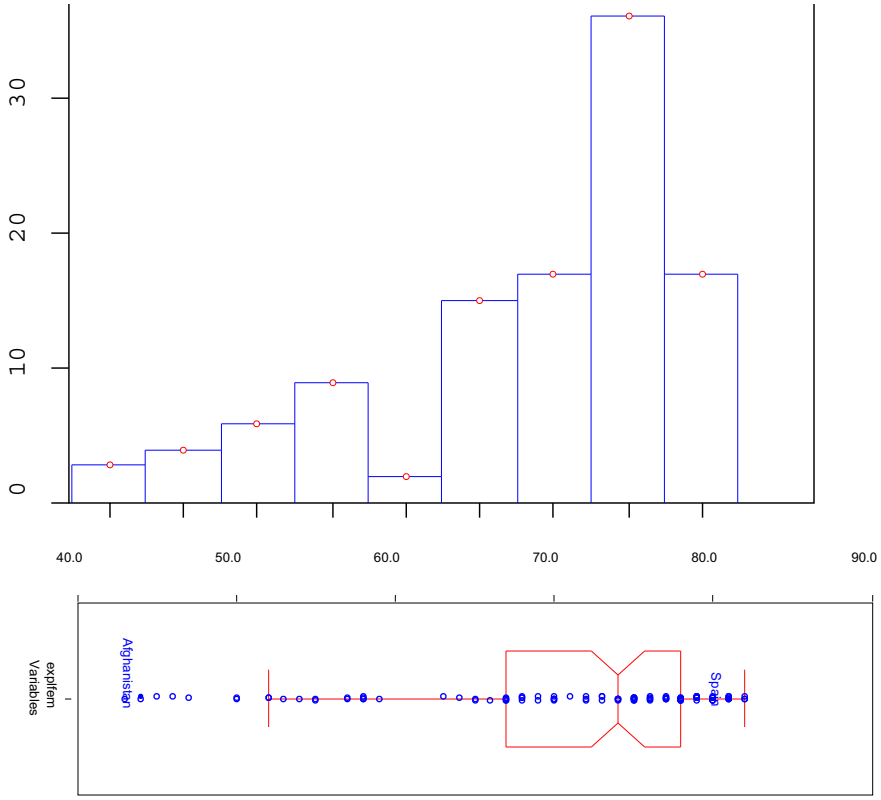
- Una versión más conocida del diagrama de puntos con añadidos es el diagramas de box and whiskers introducido por Tukey. Este gráfico representa la mediana, el rango intercuartil (y otras cosas).
- Veamos un ejemplo basado en la expectativa de vida en diferentes países en el año 95. En el gráfico siguiente se muestran la diferencia en expectativa de vida entre hombres y mujeres

– La línea central representa la mediana, los lados el cuartil primero y el tercero (y la distancia intermedia el rango intercuartil que incluye el 50% de los datos)



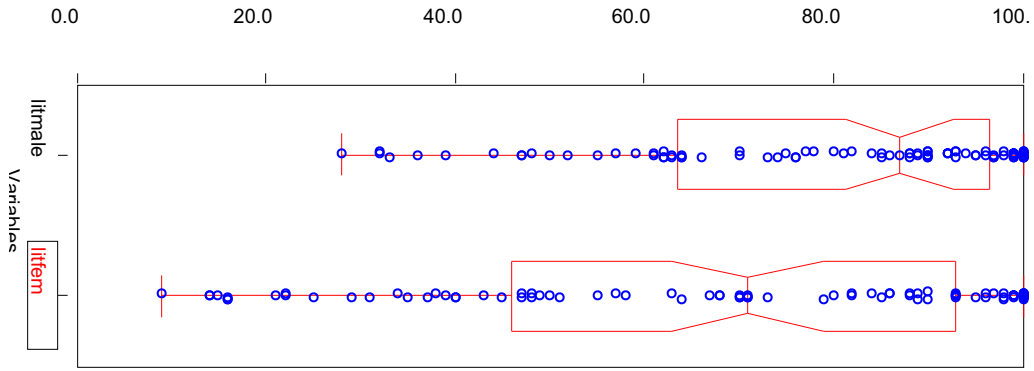
– Fijaros que el cuerpo central no tiene por qué ser simétrico (a diferencia del diamante que sí era simétrico). Eso nos da una idea de si los datos son asimétricos.

– Por ejemplo, si vemos los datos de la expectativa de vida femenina por separado con un histograma y un diagrama de cajas vemos la correspondencia



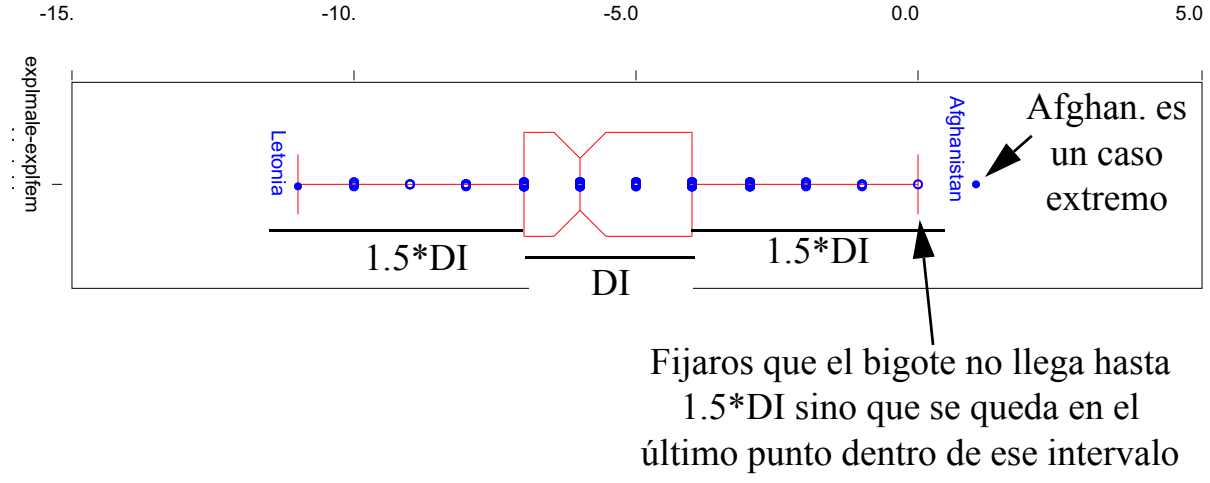
La correspondencia no es exacta pero aquí se puede ver aproximadamente como un diagrama de cajas se corresponde con el histograma

- Los diagramas de cajas son también una buena manera de comparar variables



Vemos que la mediana de la alfabetización femenina es menor que la masculina y también que la variabilidad es menor

- Los bigotes aportan información acerca de los valores extremos pero atención su definición es un poco peculiar
  - Los bigotes van hasta el último punto que se encuentra por encima de 1.5 veces la distancia intercuartil con respecto al primer cuartil o por debajo de 1.5 veces con respecto al tercer cuartil. Gráficamente se entiende mejor

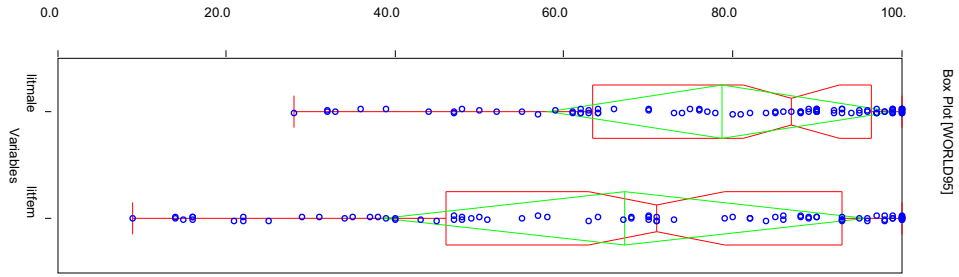


- Los puntos que están más alla del 1.5*DI se consideran valores extremos de nivel medio (Afganistan por ejemplo tiene una esperanza de vida para las mujeres menor que para los hombres lo cual es contrario a lo que ocurre en el resto del mundo) Por otro lado, los que están a más de 3*DI se consideran muuuuy extremos
- ¿A qué se debe la regla de 1.5*DI? En una ocasión le preguntaron a Tukey (el que se inventó el boxplot) por qué usar la regla del 1.5. Él contestó que 1 sería un intervalo demasiado pequeño, y 2 sería un intervalo demasiado grande

- Mi explicación es que por encima de 1.5*DI es habitual que haya uno o dos casos a menudo. Por encima de 3*DI es habitual que no haya ninguno.

3.21.Todavía más completo, diamantes más cajas

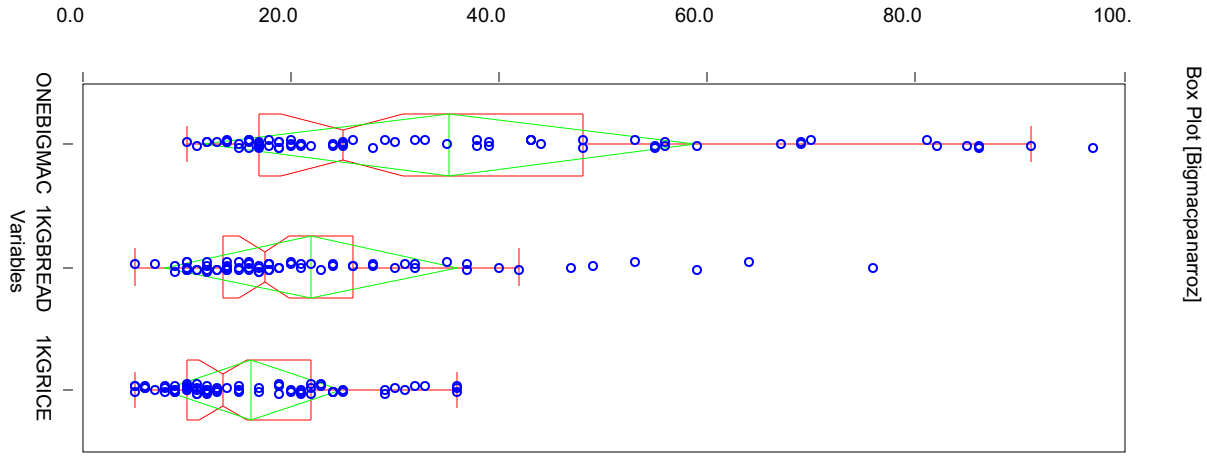
- Los diamantes están diseñados para que se puedan utilizar junto a las cajas



Fijaros como medias y medianas no coinciden. ESto indica la asimetría que hay en los datos. En este caso, las medianas son más optimistas que las medias ya que hay valores extremos en ambas variables por la parte de abajo

ACTIVIDADES

EJERCICIO 3.21.1 Comenta el siguiente gráfico de los precios de productos básicos en ciudades del mundo



3.22.Otros indicadores de variabilidad, asimetría, etc.

- Hay una serie de indicadores numéricos que complementan lo visto anteriormente y que están en el programa. Estos son:
  - De variabilidad: La amplitud, la varianza, la desviación típica poblacional y el coeficiente de variación
  - De Asimetría: El índice intercuartilico y el de Fisher

- La amplitud: Es la diferencia entre el máximo y el mínimo. Puede tener utilidad pero es muy poco robusta y depende de sólo dos valores por lo que puede ser engañosa

- La varianza: La varianza es la desviación típica al cuadrado. Es importante sobre todo a la hora de realizar cálculos pero en cuanto a darle una interpretación es más difícil de hacerlo que la desviación típica. A veces, se utiliza el término varianza en lugar de variabilidad

$$s^2 = \frac{\sum (y - \bar{y})^2}{n}$$

- Desviación típica poblacional: En ocasiones, la fórmula de la desviación típica la vereis así. Esta fórmula sirve para estimar la desviación típica de la población a partir de una muestra.

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

- Coeficiente de variación: El coeficiente de variación permite comparar datos cuando tienen diferentes medias.

$$CV_x = \frac{S_x}{X}100$$

- El índice de asimetría: Es una forma de valorar la asimetría mediante un coeficiente numérico. Yo recomiendo mirar el gráfico. Valores menores de 0

$$AS = \frac{\bar{X} - Mo}{S_x}$$

indican asimetría negativa, mayores positiva y cero simetría.

- El índice de asimetría de Fisher: mide lo mismo que el anterior. Se interpreta igual que el anterior

$$ASF = \frac{\sum (X_i - \bar{X})^3}{nS_x^{\ 3}}$$

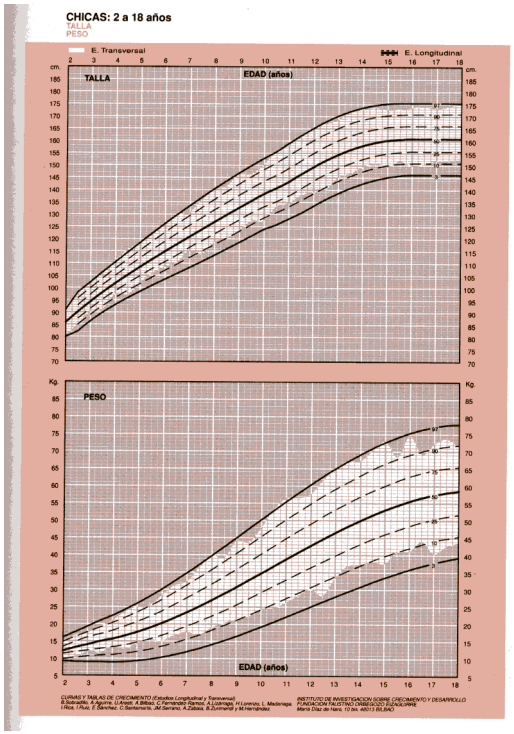
3.23.Las posiciones individuales

Ejemplo

Un ejemplo que muchos de nosotros tenemos que ver a lo largo de la vida son los datos de peso, altura, etc. de los niños. Esa información la anota el pediatra en un gráfico que aquí en la comunidad valenciana tiene la forma mostrada aquí.En ese gráfico tenemos los percentiles en la parte de arriba de las líneas que indican los márgenes normales de peso y altura. Mirando esas líneas podemos convertir los resultados en una escala de tantos por ciento, lo cual resulta fácil de comunicar ya que la mayoría de la gente

entiende esa escala. Además, resulta fácil comparar la evalución de una niña. Si un año, una niña está en el percentil 10 y al año siguiente en el 50 decimos que ha ganado peso. Fijaros que una niña puede ganar peso de una edad a otra y sin embargo, en percen-

tiles, haber perdido peso.



- Hasta ahora hemos visto que para describir un conjunto de datos nos interesaba ver la tendencia central, la variación y un número de otros pequeños detalles.

- En esta sección veremos como describir las posiciones de cada uno de los casos. Es decir, si un valor determinado es alto, es bajo o no lo es, y cual es la magnitud en la que esto ocurre. Los dos problemas fundamentales que hay que resolver a este respecto son:
  - La comunicación: Una puntuación en una escala debe ser sencilla de entender
  - La comparación: Hay que hacer comparaciones entre diferentes puntuaciones y es necesario que la escala admita esa comparación inmediatamente.
- Los métodos para indicar posiciones son:
  - Los percentiles
  - Las puntuaciones típicas

– Percentiles normalizados (teniéndolo todo)

3.24.Percentiles

- La idea de los percentiles consiste en calcular el tanto por ciento del total que una puntuación tiene por debajo
  - Nos da unos valores teóricos entre 100 y 0 (teóricos porque es raro tener un 100 o un 0 en esa escala)
  - Nos permite hacer comparaciones entre variables que son interpretables (un percentil 10 a los 5 años y un percentil 20 a los 6 años para un mismo niño sugiere un aumento de peso mientras que pasar de 11 a 14 kilos no se sabe si es aumento o no)
- Dos conceptos complementarios son los de Percentiles y Rango Percentil. Estos conceptos son complementarios y es muy fácil confundirlos

  - El percentil es la puntuación que deja por debajo de sí un porcentaje de casos (p.e. percentil 10 para el peso a los 12 años es 30 kgs de peso. Percentil (10)=30)
  - El rango percentil es el tanto por ciento que deja por debajo una puntuación dada (p.e. rango percentil de 30 kgs de peso a los 12 años es 10. RP(30)=10)
- Es necesario aprender a hacer las dos operaciones en la práctica. Por ejemplo, con la altura y el peso de las niñas
  - Nos pueden decir un rango percentil y una edad y queremos saber qué peso o altura significa eso.

- Nos pueden decir un peso o una altura y queremos saber qué porcentaje de casos hay por debajo de esos valores (es decir cual es el rango percentil)

ACTIVIDADES

EJERCICIO 3.24.1 Calcula en el gráfico de talla y peso el rango percentil de una niña de 10 años con 40 kilos de peso y que mide 1.50 de altura

EJERCICIO 3.24.2 ¿Es posible que una niña de 12 años pese 70 kgs?

EJERCICIO 3.24.3 ¿Cuál es la mediana de peso a los 3 años de edad?

EJERCICIO 3.24.4 ¿Cuál es el percentil 10 en peso a los 18 años de edad?

EJERCICIO 3.24.5 ¿Cuál es el rango percentil de una niña de 18 años con 60 kilos?

EJERCICIO 3.24.6 Si digo algo en kilos...¿qué es lo que puedo pedir? ¿Por qué? ¿En qué escala estará la respuesta?

EJERCICIO 3.24.7 Si digo algo en porcentajes...¿qué es lo que voy a pedir? ¿En qué escala estará la respuesta? ¿Por qué?

3.25.¿Cómo se calculan los rangos percentiles y los percentiles?

- Para hacer estos cálculos se necesita saber convertir unos datos en rangos.
  - Convertir en rangos sería como numerar del mayor al menor si no hubiera empates. Por ejemplo, la expectativa de vida femenina de Afganistán, Haití y Camboya es respectivamente de 44, 47 y 52. Los rangos de estos tres países serían 1, 2 y 3.

– Cuando hay empates lo que se hace es poner el punto medio a las posiciones que les tocarían a los datos si no hubiera empates. Por ejemplo, los peores 9 países en expectativa de vida masculina son los siguientes.

País	ExpcVida-Mas	Rango
Uganda	41	2.000
Rep. C. Afri	41	2.000
Tanzania	41	2.000
Ruanda	43	4.500
Haití	43	4.500
Zambia	44	6.000
Afganistán	45	7.000
Burundi	46	8.000
Burkina Faso	47	9.000

- Como los tres primeros países están empatados se les pone un 2 tras hacer (1+2+3)/3=2. El cuarto y el quinto se les pone 4.5 y el resto como no hay empates sigue la cuenta.

- En resumen, para calcular los rangos percentiles hacemos:
  - Convertir los datos a rangos
  - Dividir el rango del valor que nos interesa por el total de casos
  - Multiplicar por 100

Ejemplo

En los datos de expectativa de vida que aparecen en el SPSS sería interesante saber cual es el rango percentil que ocupa España. En esos datos hay 109 países. La expectativa de vida en España femenina en el año 95 era de 81 años y la masculina era de 74 años. España tiene el rango 103 en expectativa de vida femenina así que (103/109)*100=94.5. En expectativa de vida masculina estamos en 96.5 así que (96.5/109)*100=88.53. ¿Qué podríamos decir de estos dos resultados?

- Para calcular los Percentiles hacemos
  - Convertir los datos a rangos
  - Calcula el número de casos que supone un porcentaje dado del total de casos (p.e. si tienes 150 casos y quieres un 10% el número de casos es 15)
  - Al valor anterior súmale 0.5 y redondea al entero más próximo. Este es el rango del percentil que buscas.
  - Toma el valor en los datos que corresponda al rango calculado en el paso anterior.

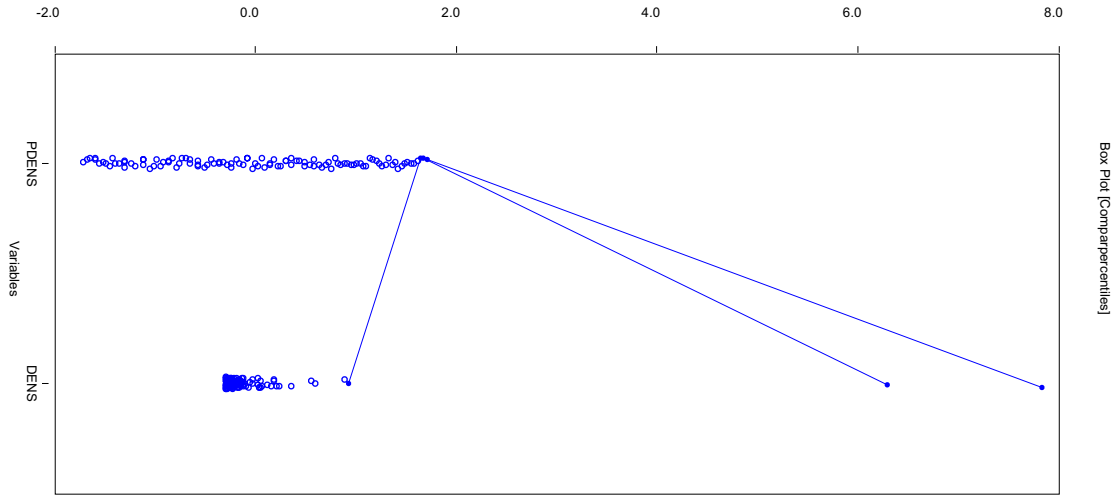
Ejemplo

Vamos a calcular la expectativa de vida femenina que corresponde a tener un percentil 80 (es decir, el 80% de los países tendrán una expectativa de vida femenina más baja). Teníamos 109 países, así que el 80% es (80*109)/ 100=87.2. A ese valor le sumamos 0.5 y redondeamos y sale 88. No obstante, como hay empates, no hay un rango de 88 así que cogemos el superior que es 89. Ese valor corresponde con la puntuación de 79.

3.26.Inconvenientes de los rangos percentiles

- Usar percentiles presenta el inconveniente de que la información de las distancias entre los puntos se pierde y sólo queda la información del porcentaje. Así, un punto porcentual puede significar una gran distancia en los valores de la escala original mientras que en otras ocasiones puede no significar mucho.

- Visto gráficamente (densidad en países del mundo).

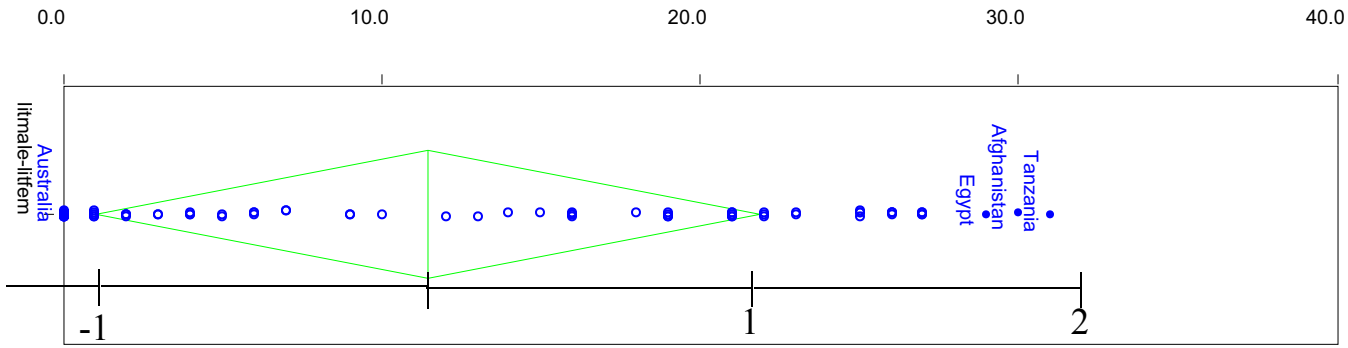


  - Las diferencias entre los tres últimos países son mucho más grandes y diferentes entre ellas (hay más diferencia entre el tercero y el segundo que entre el primero y el segundo), en rangos percentiles esa información desaparece.

  - En rangos percentiles, ***sólo sabemos de dos países consecutivos que un país está por encima del otro pero no podemos ver la distancia que hay entre ellos.***
- En conclusión, convertir a rangos percentiles una variable nos hace perder información
- Una alternativa que no tiene esos problemas está en la siguiente sección

3.27.La desviación típica como una regla

- Recordar el gráfico de diamantes para la diferencia en alfabetización entre hombres y mujeres



La desviación típica es como una regla que se aplica para medir las distancias respecto del centro

  - Cada puntuación puede medirse con respecto a su media en la unidad de medida “desviaciones tipicas”

– Esto nos permite decir por ejemplo: “este país está 1 desviación típica por encima de la media” o “este país está a media desviación típica por debajo de la media”

3.28.Cálculo

- A las diferencias respecto de la media en términos de desviaciones típicas las llamamos puntuaciones típicas y se calculan así:

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

– En la siguiente tabla teneis una serie de puntuaciones de la alfabetización de mujeres, y su puntuación típica equivalente. Para hacer este cálculo hace falta la media (67.26) y la d.t.(28.61) de la alfabetización femenina

Afghanistan	14.00	-1.86
Saudi Arabia	48.00	-0.67
Argentina	95.00	0.97
Armenia	100.00	1.14
Australia	100.00	1.14
Azerbaijan	100.00	1.14
Bahrein	55.00	-0.43
Bangladesh	22.00	-1.58
Barbados	99.00	1.11
Bielorussia	100.00	1.14
Bolivia	71.00	0.13
Botswana	16.00	-1.79
Brasil	80.00	0.45
Burkina Faso	9.00	-2.04
Burundi	40.00	-0.95

– Por ejemplo para calcular el primero de los datos hacemos

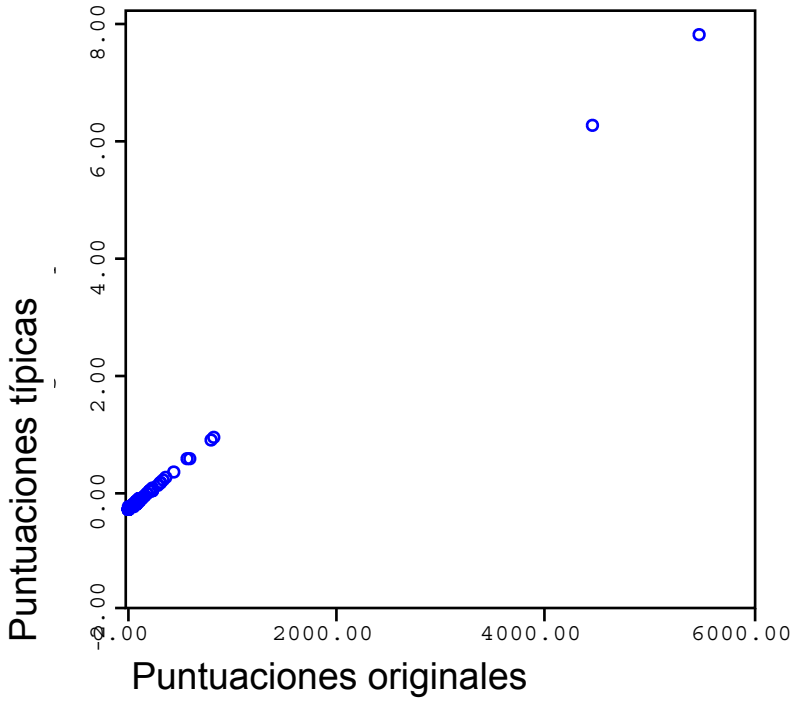
$$z_{Afganis\,\tan} = \frac{(14 - 67.26)}{28.61}$$

ACTIVIDADES

EJERCICIO 3.28.1 Calcular el resto de las puntuaciones típicas de la tabla anterior

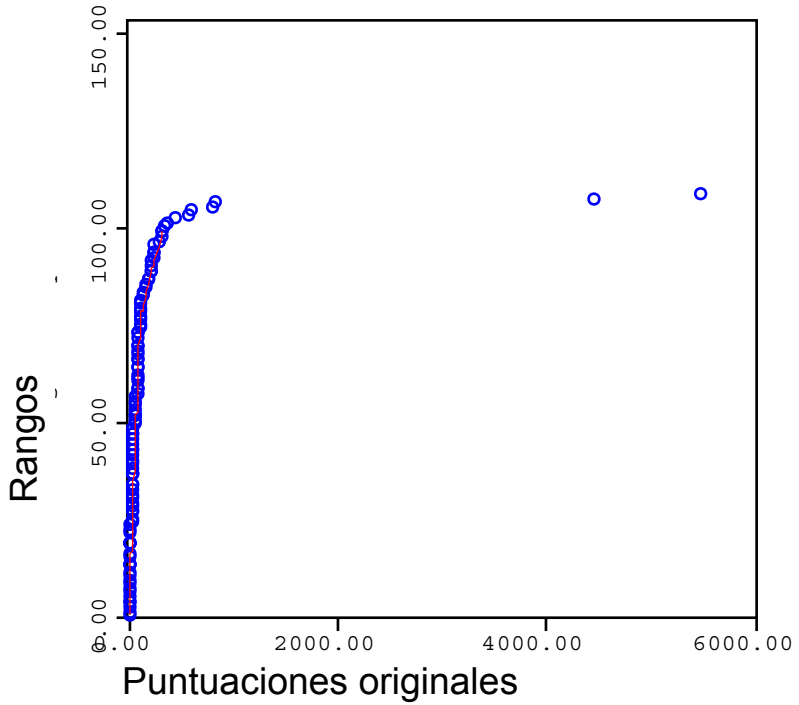
3.29.Propiedades de las puntuaciones típicas

- Las puntuaciones típicas están relacionadas linealmente con las puntuaciones originales



– Esto significa que las distancias que había originalmente entre los puntos se respetan proporcionalmente.

- Podemos comparar eso con el efecto que produce pasar a rangos (que es igual a rangos percentiles)



  - En este segundo caso, la transformación tuerce la relación entre la variable original y la transformada.

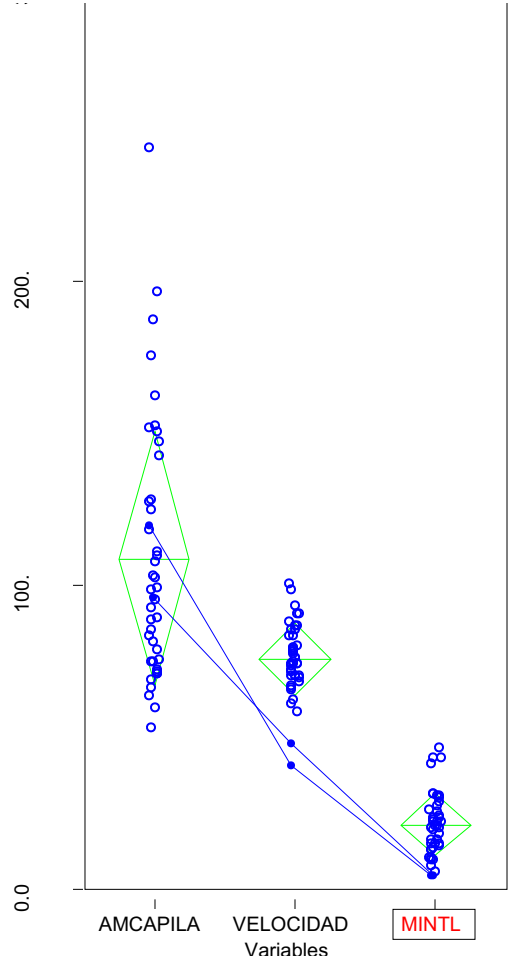
- Nota: el primer gráfico es un ejemplo de una transformación lineal. Una transformación lineal es la que hacemos cuando cambiamos entre escalas que son equivalentes como por ejemplo de kilos a libras, de grados Celsius a Farenheit, o de monedas. El segundo gráfico es un ejemplo de transformación no lineal.
  - Una transformación lineal es la consecuencia de sumar o restar un valor a todas las puntuaciones o de multiplicar o dividir un valor a todas las puntuaciones
  - Cuando sumamos o restamos una constante a todas las puntuaciones las medidas de tendencia central suben o bajan en ese valor (pero las medidas de variación no cambian). Por ejemplo, si a los resultados de un exámen les sumo un punto a todos

los alumnos, la media sube un punto pero las distancias entre el primero y el último seguirán siendo las mismas.

– Cuando multiplicamos o dividimos todas las puntuaciones por una constante tanto las medidas de tendencia central como las de variación son multiplicadas o divididas por ese valor

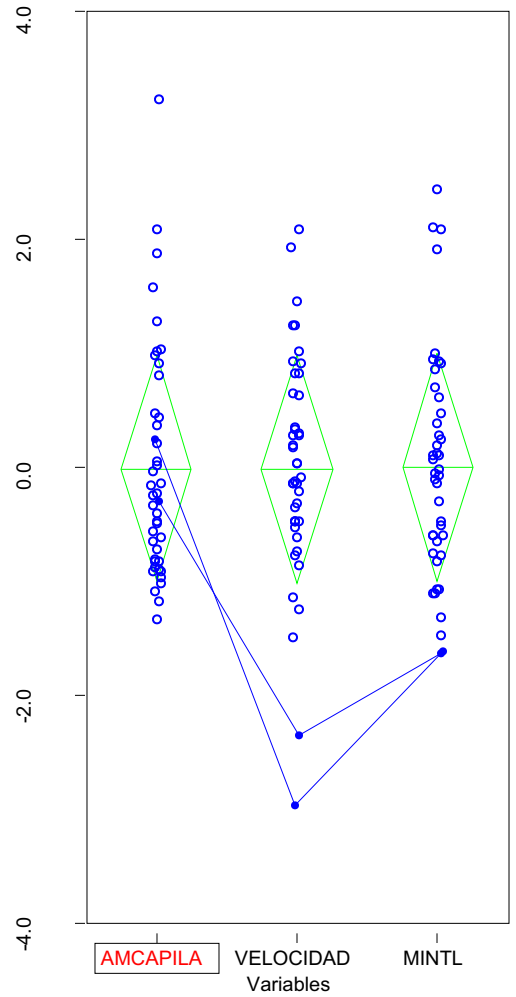
- Otras propiedades de las puntuaciones típicas:
  - La media de las puntuaciones típicas es cero (lógico, porque el primer paso consiste en restar la media de todas las puntuaciones)
  - La desviación típica es 1 (lógico, porque se divide todas las puntuaciones originales por la desviación típica)
- En resumen, cambiar a puntuaciones típicas tiene las siguientes consecuencias
  - Cambia el centro de los datos y lo pone en el cero
  - Cambia la variación de los datos y lo convierte en 1
  - NO cambia la forma de los datos (un hist. de los datos originales y de las p. típicas la forma es similar)

3.30.Comparación de variables con puntuaciones típicas



- En el gráfico anterior, las comparaciones entre variables para cada sujeto eran difíciles porque los datos estaban en diferentes escalas

- La forma de solucionar eso es utilizar puntuaciones típicas

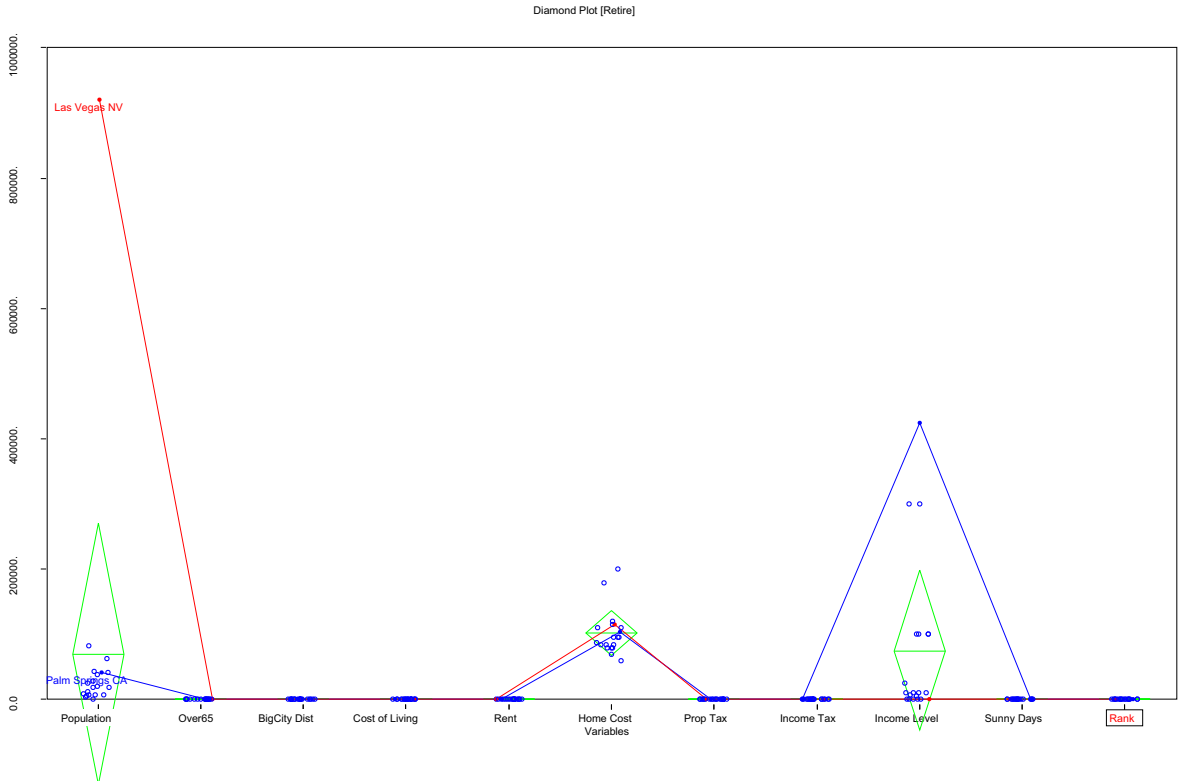


- Al usar puntuaciones típicas, las medias y las desviaciones típicas son iguales (los diamantes son iguales)
- Las puntuaciones individuales son más fáciles de valorar (vemos que los dos sujetos marcados son especialmente lentos aunque con un bajo Mintl que indica que condujeron bien)

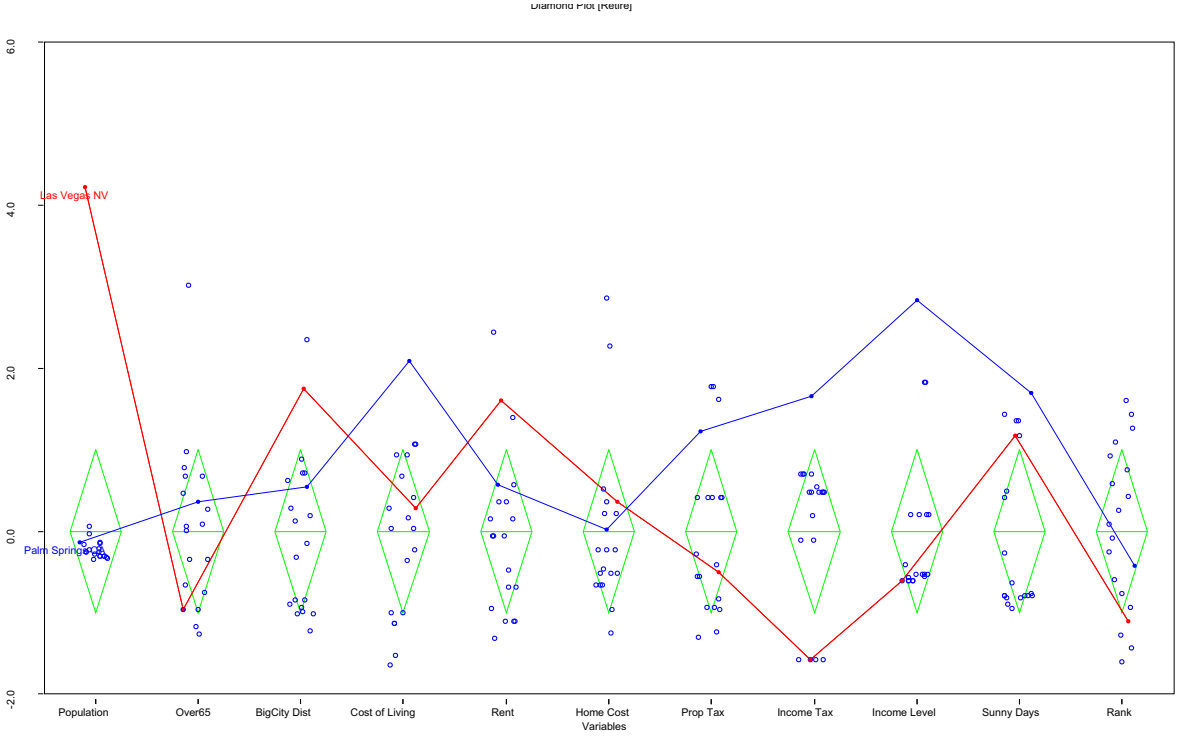
Ejemplo

Un ejemplo más radical puede ayudarnos a entender la ventaja de las puntuaciones típicas. En Estados Unidos es normal que al jubilarse muchos opten por ir a vivir a un lugar especial. En una revista se indicaron una serie de características que pueden ser atractivas para elegir la mejor ciudad. Son cosas del tipo dias soleados, impuestos, coste de la vida, coste de una casa, etc. En los gráficos he señalado dos ciudades que me han llamado la atención: Las Vegas en Nevada, y Palm Springs en California

– Sin normalizar



– Normalizado



3.31.Percentiles normalizados

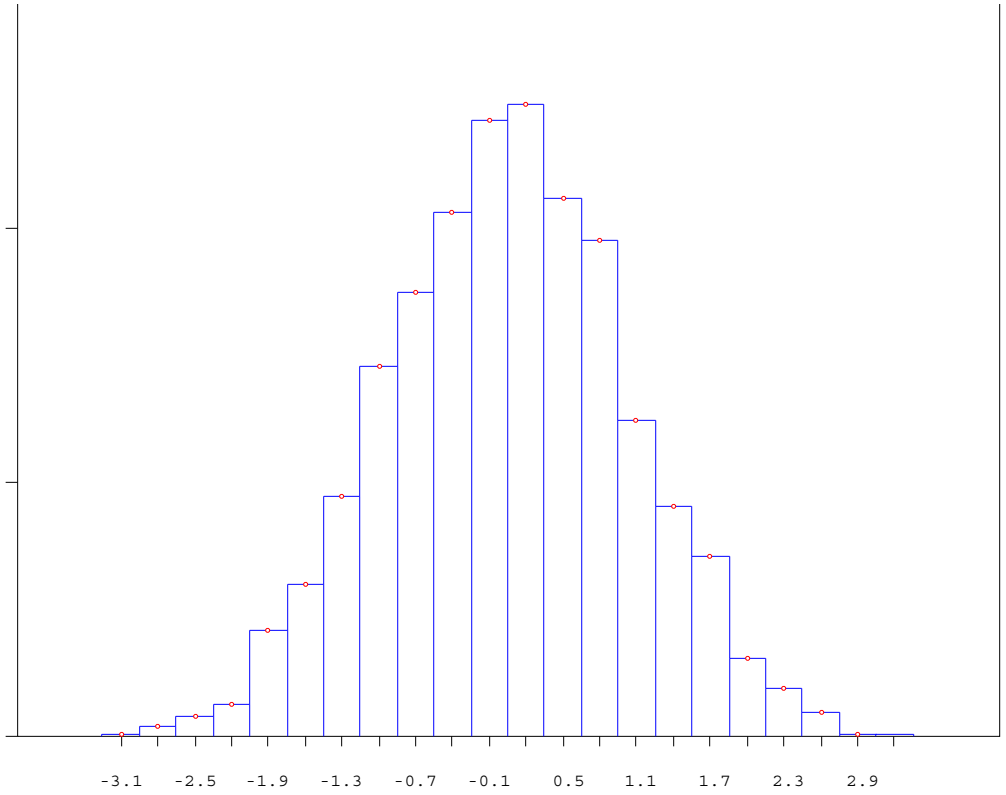
- En las secciones anteriores hemos visto que_
  - Los percentiles son fáciles de interpretar pero distorsionan la forma de los datos
  - Las puntuaciones típicas no distorsionan la forma de los datos pero no son fáciles de interpretar ya que no tenemos una referencia de cómo de inusual es una puntuación típica dada
- ¿Es posible tenerlo todo?->Percentiles normalizados

- La idea de los percentiles normalizados es obtener los porcentajes que quedan por debajo de una puntuación a partir de un modelo teórico desarrollado por los matemáticos
- Ese modelo matemático se denomina el modelo Normal. Veremos ese modelo en primer lugar y luego pasaremos a ver como podemos utilizarlo para calcular los percentiles normalizados

3.32.El modelo de distribución Normal de los datos

- En secciones anteriores hemos visto que la forma de la distribución de unos datos puede variar según el problema concreto
  - Para datos económicos es habitual que la distribución sea asimétrica positiva
  - En otras ocasiones la asimetría es negativa
  - Otra posibilidad es la de los juegos de azar, en ese caso, la distribución es uniforme. Por ejemplo, si lanzamos un 120000 veces un dado, nos saldría cada valor del dado unas 20000 veces

- Un modelo de distribución de datos de gran importancia es la distribución Normal. Esta distribución es simétrica, con un centro en los datos y dos colas que se extienden hacia la derecha e izquierda. Un ejemplo de datos que seguirían la distribución normal muy idealizado sería:

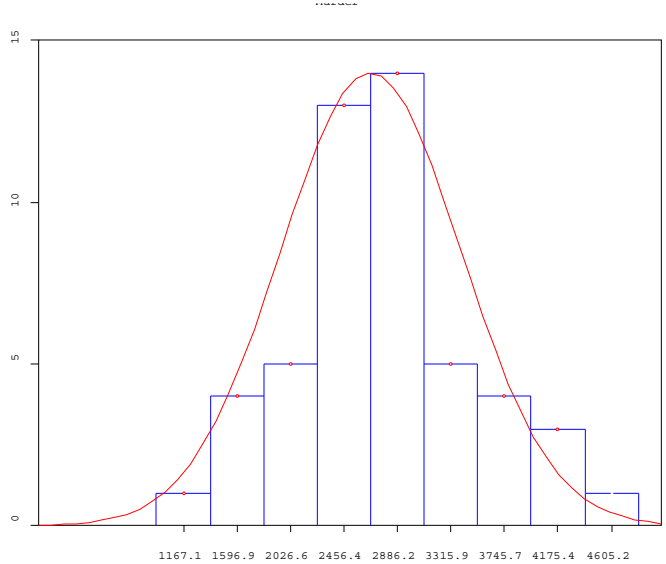


- Existe una gran variedad de situaciones en las que cuando sacamos los datos y los representamos nos aparece una distribución de este tipo

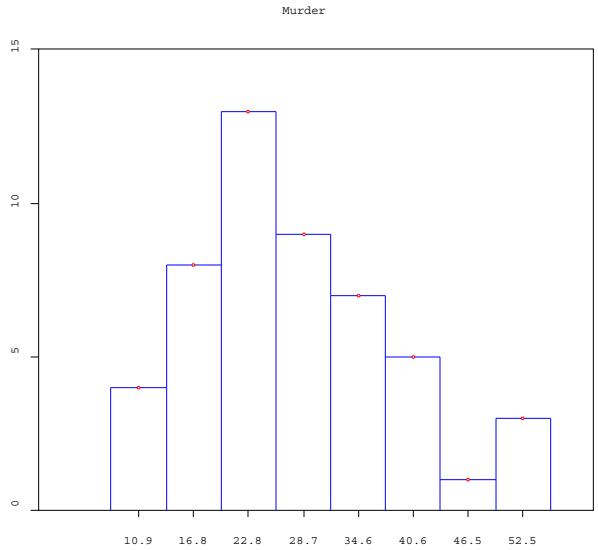
Ejemplo

Mostraremos los datos de crímenes por 100.000 habitantes en cada uno de los 50 estados americanos en el año 1980. Hay varios tipos de crímenes.Fijaros que la línea roja es una distribución normal idealizada y es la de los datos que representamos.

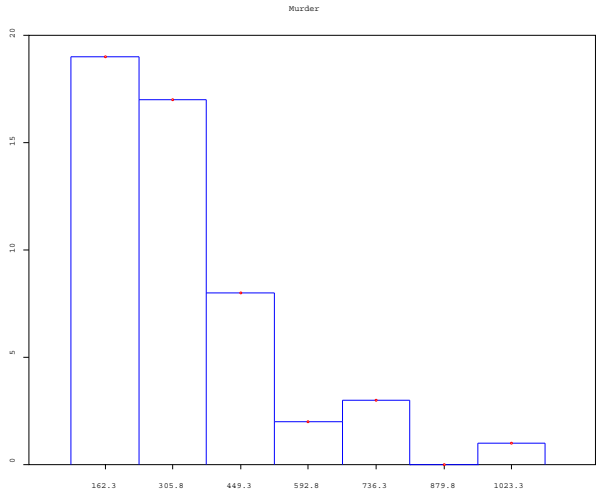
– Para el hurto vemos que la distribución se parece bastante a la normal



– En cambio las violaciones tienen una ligera asimetría derecha

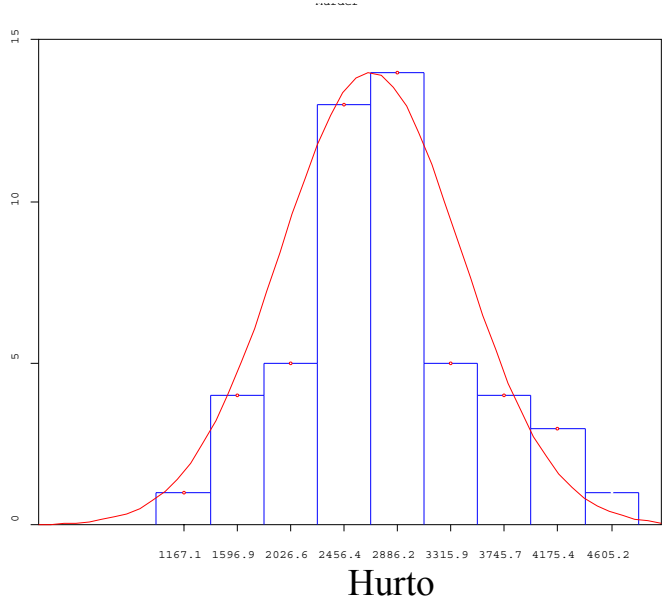


– Finalmente, el robo de coches parece asimétrico derecho y no se ajusta a la distribución normal



Robo Coches

- La distribución normal puede ser descrita de una manera precisa mediante una fórmula matemática. Esa fórmula es la que nos permite dibujar la línea roja que hemos puesto en el primer gráfico (la fórmula no suele poners en cursos introductorios)



  - Los datos reales no se van a ajustar exactamente a esa curva nunca, sólo aproximadamente

– A veces, decir que unos datos siguen una distribución normal es más una cuestión de creencias que de pruebas empíricas, uno puede mantener que unos datos deben seguir la distribución normal a pesar de que los histogramas muestren lo contrario (por ejemplo, un grupo de estudiantes en un colegio dado pueden seguir una distribución asimétrica en inteligencia y uno puede seguir manteniendo que la inteligencia se distribuye normalmente y que ese modelo es válido para esos estudiantes)

- ¿De dónde viene la distribución normal? ¿Por qué es tan habitual que haya situaciones en la vida real en que los datos siguen la distribución normal?
  - La distribución normal surje de sumar una serie de variables aleatorias distribuidas de modo uniforme
  - Eso quiere decir, que cualquier cosa que sea la suma de una serie de factores individuales independientes que pueden variar de una manera impredecible puede acabar distribuyéndose de manera normal. Cada uno de esos factores puede ser desconocido, pero la idea es que el efecto de muchos de ellos combinados producirán algo que se distribuye normalmente

- Ejemplos de medidas que suelen considerarse normales son:
  - Algunas medidas biológicas (medidas de las uñas, garras, pelo, dientes, presión sanguínea en adultos). Otras medidas fisiológicas pueden seguir esa distribución pero no hay razón para asumirlo.
  - Errores de medida: Si uno mide lo mismo varias veces con un aparato que tiene cierto error de medida es típico que cada vez tengamos una medida ligeramente diferente. Esas desviaciones se supone que se distribuyen normalmente
  - Los resultados de tests suelen ser construídos de tal manera que el resultado se distribuye normalmente

- Ejemplos de medidas que NO son normales
  - Las variables financieras NO suelen seguir la distribución normal (sin embargo, el logaritmo de esas variables sí que son normales)
  - El tamaño de los animales adultos NO sigue la distribución normal (pero el logaritmo sí)

3.33.¿Qué utilidad tiene la distribución normal?

- Tener una descripción mátemática de una distribución de probabilidad nos permite saber qué resultados podemos esperar y cuándo esos resultados son inesperados.

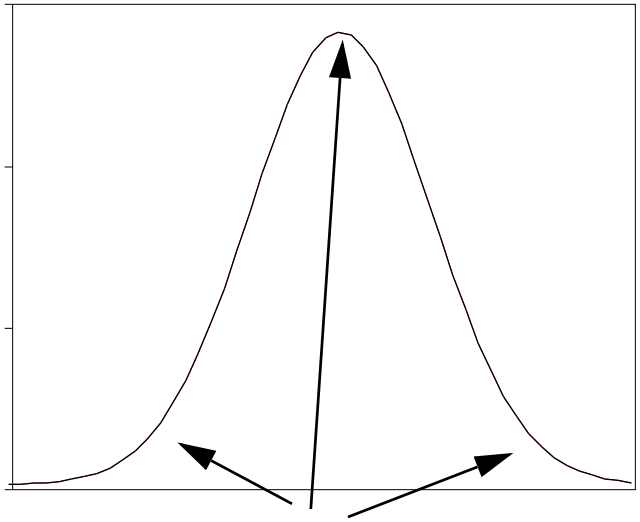
- Pongamos por ejemplo la distribución uniforme aplicada a los lanzamientos de un dado.
  - Tiramos un dado 600 veces y los resultados son los siguientes. ¿Diriais que hay un problema?

Tabla 11: Resultados de lanzar un dado 600 veces

1	2	3	4	5	6
101	99	102	98	10	190

  - Obviamente, el valor 5 ha salido muy pocas veces y el 6 ha salido casi el doble de lo que esperaríamos. Aquí hay algo extraño (que deberíamos investigar)
  - Fijaros que sabemos esto por que conocemos la probabilidad teórica con la que tendrían que salir los resultados del dado (1/6)

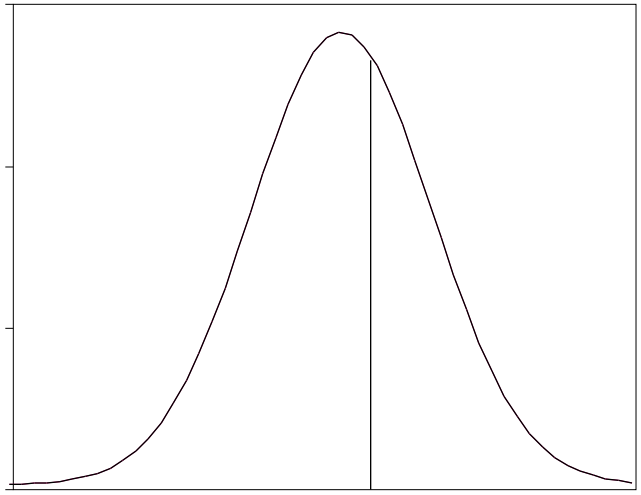
- Las probabilidades teóricas de que salgan ciertos resultados para datos que siguen la distribución normal se suelen representar mediante la curva que ya hemos visto varias veces. Por ejemplo,



Cuanto más alto, más veces salen valores de ese tipo. En la curva normal, salen más veces los valores medios, y menos los extremos

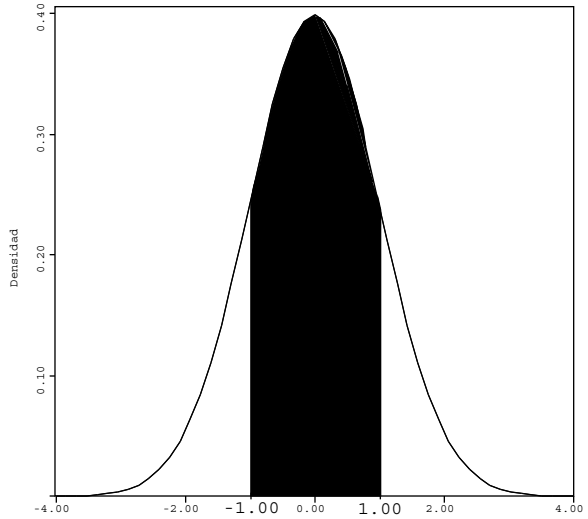
- Para saber exactamente las probabilidades de un resultado con la distribución normal tenemos:
  - Tablas
  - Ordenadores
  - Calculadoras
  - Memoria (este es el método que yo recomiendo)

- Algunos valores interesantes para memorizar
  - Probabilidad de un valor concreto=0. Con la distribución normal siempre hay que usar intervalos



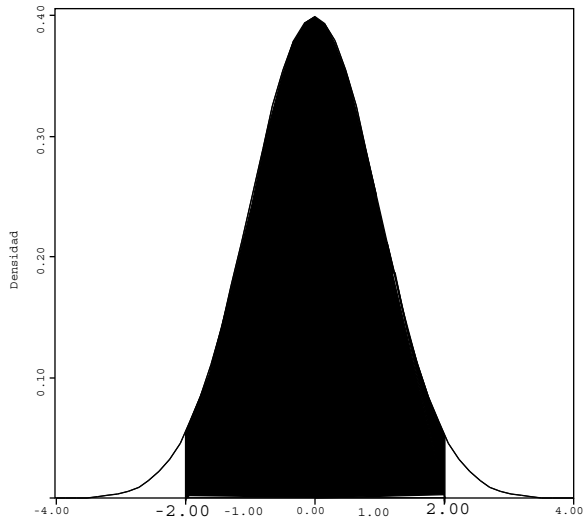
La distribución normal es contínua (valen todos los decímales) así que una probabilidad puntual es cero

– Probabilidad dentro del intervalo una desviación típica por arriba o por debajo de la media (es decir, de ser del montón)->0.68 (el 68% de los datos están entre -1 y +1 desviaciones típicas de la media)



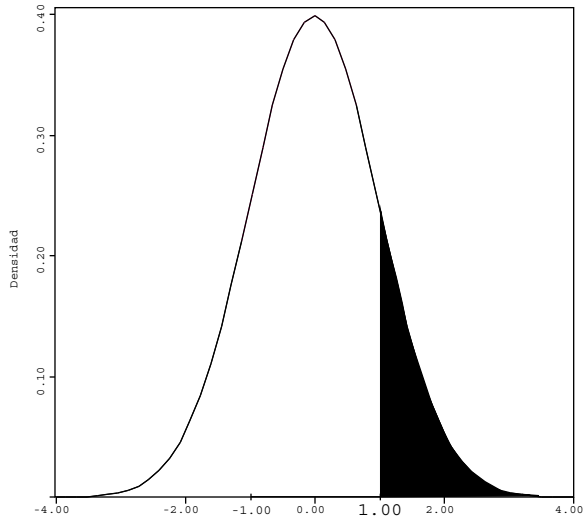
– Como el área total es la probabilidad de estar fuera del intervalo anterior es 1-0.68=0.32 (el 32% fuera)

– ¿Y dos desviaciones tipicas? 0.9545 (el 95.5% de los datos están entre dos desviaciones típicas)

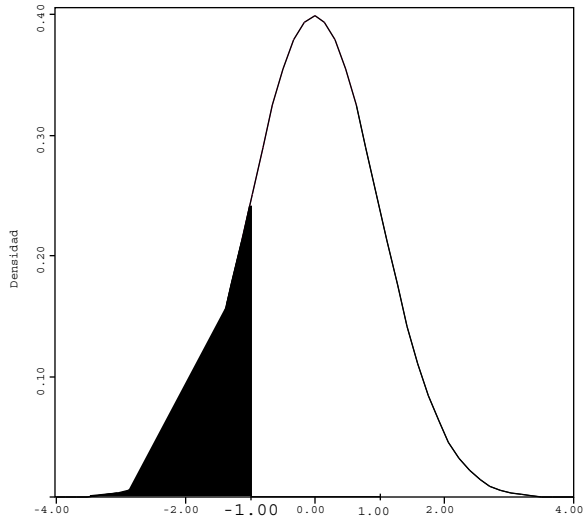


– ¿Y fuera? 1-0.9545=0.045 (el 4.5% está fuera)
– Si damos el porcentaje de casos que están por debajo de un valor dado estamos usando la curva normal como hacíamos cuando trabajábamos con

rangos percentiles. Por ejemplo, si una persona está 1 desviación típica por encima de la media el porcentaje que tiene por debajo es del 84% y el 16% por encima



– Si está una desviación típica por debajo entonces es al revés. El 84% está por encima y el 16% por debajo.



– ¿Dos desviaciones típicas por encima? El 97.7% está por debajo.

- ¿Y tres desviaciones típicas por encima? Entonces, por debajo está el 99.8%
- ¿Y cuatro? Entonces por debajo está el 99.9%

- El problema contrario al que estamos viendo también tiene sentido. En concreto, es bastante habitual tener que buscar:
  - Qué dos puntuaciones típicas dejan entre sí el 95% del area debajo de la curva normal-> -1.96 y 1.96 (estos valores están redondeados)
  - Qué dos puntuaciones típicas dejan entre sí el 99% del area debajo de la curva normal->-2.56 y 2.56 (estos valores están redondeados)

- En resúmen,
  - Previo a calcular percentiles normalizados hay que plantearse si es razonable pensar que los datos que usamos siguen la distribución normal. Hacer un gráfico y pensar un poco sobre el tema puede ayudar pero a menudo tendremos que fijarnos en lo que otros han hecho en el pasado (y confiar que sepan lo que hacen)
  - Si tenemos unas puntuaciones directas podemos cambiarlas a típicas sin perder información importante

- Si asumimos que la distribución de los datos es normal, entonces podemos usar el modelo teórico de la distribución normal para calcular percentiles (percentiles normalizados)
- NO podemos calcular la probabilidad o porcentaje de sacar una puntuación típica exacta, sólo la probabilidad de estar por debajo de esa puntuación, o por encima, o entre dos puntuaciones
- Las probabilidades son más grandes para puntuaciones típicas cercanas a cero y entre 1 y -1
- Entre las puntuaciones típicas de -2 y +2 están practicamente todas las puntuaciones.

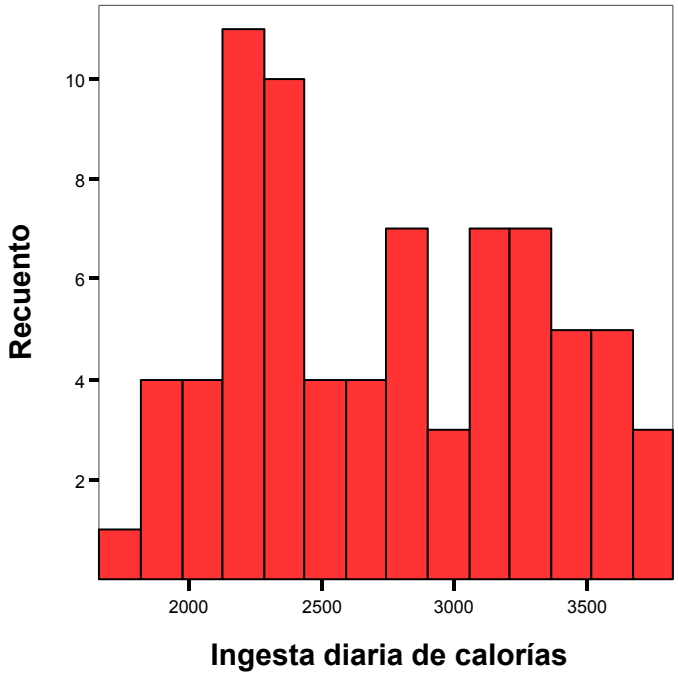
– Pasar de 2 a 3 desviaciones típicas no cambia mucho los percentiles, y de 3 a 4 todavía menos (por eso los lados de la curva normal son tan pequeños)

- ¿Y las tablas? ¿Por qué no enseñas las tablas?
  - Esta es una cita directa del libro que utilizo para preparar las clases (Stats: Data & Models, De Veaux, Velleman and Bock). “Hoy en día, encontrar percentiles en una tabla de probabilidad normal es un método de isla desierta-algo que podríamos hacer si necesitaramos desesperadamente un percentil normal y nos encontráramos atascados a kilómetros de distancia de tierra firme y con sólo una tabla de probabilidades normales (naturalmente, vosotros os podeis sentir así durante un exámen de estadística, así que es una buena idea aprender a usar estas tablas). Afortunadamente, en la mayoría de los casos podemos usar una calculadora o un ordenador”

- Buscando en Google, he encontrado la página siguiente para hacer los cálculos que necesitamos (hay muchísimas más):
- http://davidmlane.com/hyperstat/z_table.html

ACTIVIDADES

EJERCICIO 3.33.1 A continuación tienes el histograma del consumo de calorías por habitante en los países del mundo de los datos de Mundo95 ¿Es razonable considerarlo que esta variable se distribuye normalmente?



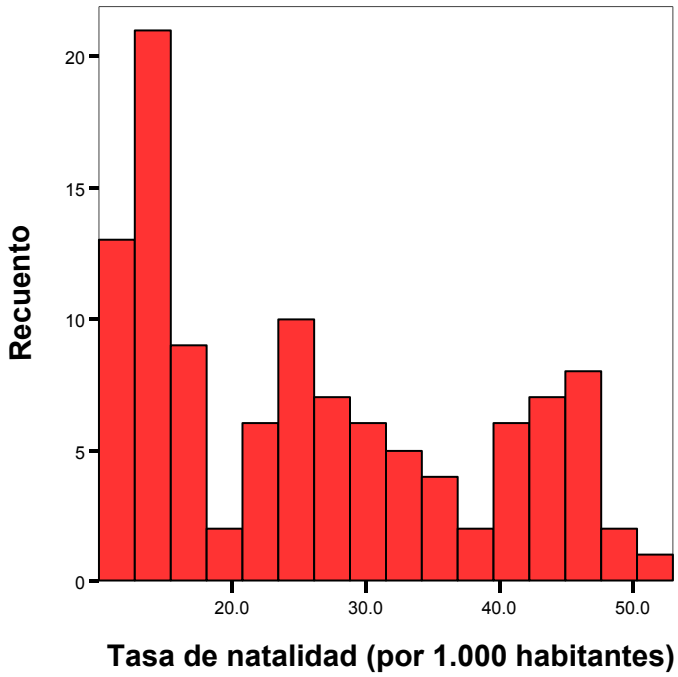
EJERCICIO 3.33.2 El consumo por habitante de calorias en España en los datos de Mundo95 es de 3572. La media de consumo de calorías para todos los países es de 2753.83 y la desviación típica es de 567.828. ¿Qué puntuación típica le corresponde a España? Cómo podrías valorar esa puntuación típica (utilizando percentiles)

EJERCICIO 3.33.3 En Somalia, el consumo de calorías por habitantes es de 1906. Valora esa puntuación.

EJERCICIO 3.33.4 En Indonesia, el consumo por habitante es de 2750. Valora ese resultado ***sin hacer ningún cálculo.***

EJERCICIO 3.33.5 La tasa de natalidad en Afganistan en puntuaciones típicas en los datos de Mundo95 es de 2.19. Valora esa puntuación.

EJERCICIO 3.33.6 A continuación se muestra el histograma de la tasa de natalidad por 1000 habitantes. ¿Dirias que el modelo normal es razonable? ¿En qué afecta eso a la interpretación de los percentiles normalizados?



EJERCICIO 3.33.7 En los datos de Mundo95 de casos de sida totales en los distintos países del mundo aparece que Estados Unidos tiene una puntuación típica de 9. ¿Podrias interpretar esa puntuación? ¿Qué significado tiene?

Parte IV
Explorando y representando datos con dos variables numéricas

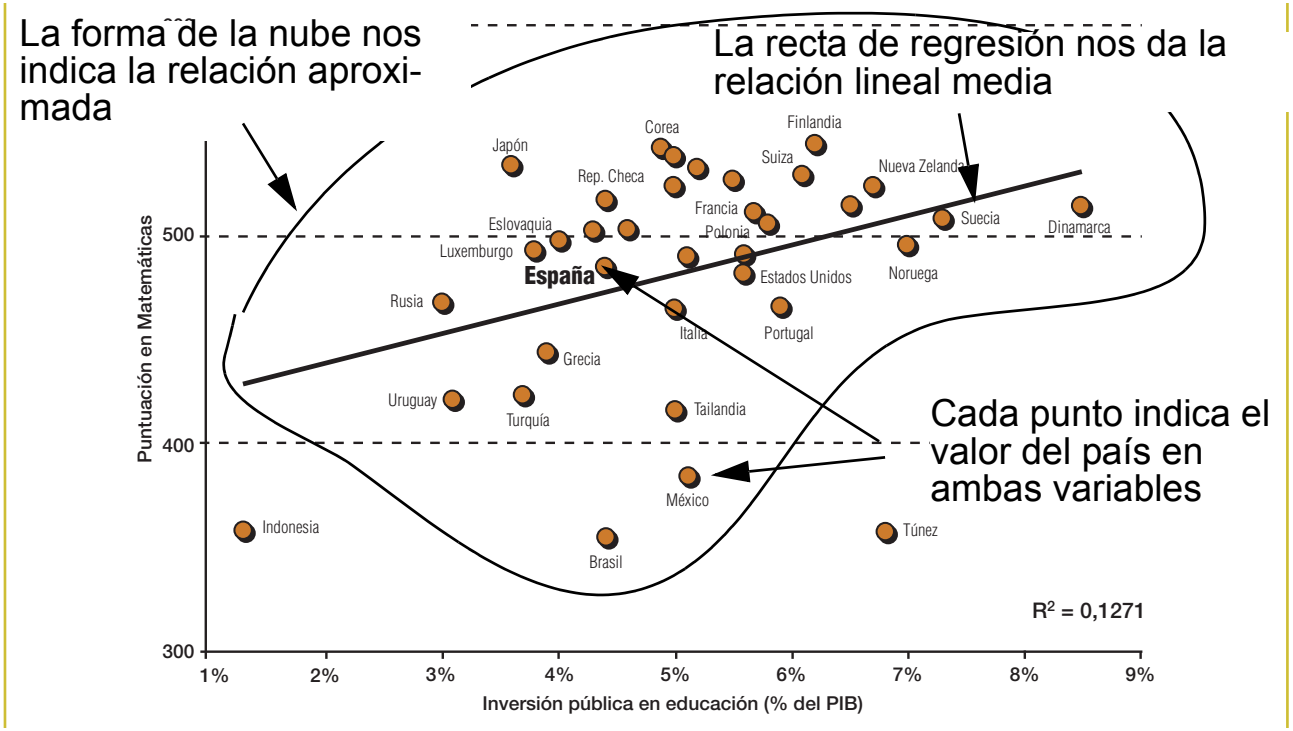
4.1.Introducción

- Hasta ahora hemos estado trabajando con una variable cada vez
- No obstante veces veíamos varias variables juntas, pero sin hacer énfasis en las ***relaciones*** entre las variables
  - Para ver las relaciones y describirlas numéricamente necesitamos nuevas herramientas que veremos en esta sección
- En esta sección veremos:
  - Como representar gráficamente dos variables y ***la relación*** entre ambas
  - Como describir numéricamente la relación entre las variables

4.2.Asociación entre dos variables contínuas: El diagrama de dispersión

Aproximación gráfica

- En el Informe Pisa se analiza la relación entre la puntuación media en matemáticas y el PIB invertido



ACTIVIDADES

EJERCICIO 4.2.1 En el informe PISA, qué puedes decir de la relación entre puntuación en Matemáticas y PIB per capita a partir del diagrama de dispersión. ¿Qué paises se ajustan peor a la relación?

EJERCICIO 4.2.2 En el informe PISA, qué puedes decir de la relación entre el índice de status socieconómico y cultura y puntuación en Matemáticas a partir del diagrama de dispersión. ¿Qué paises se ajustan peor a la relación?

EJERCICIO 4.2.3 En el informe PISA, ¿dirías que la relación entre el índice de status socieconómico y cultura y puntuación en Matemáticas podría ser curvilínea?

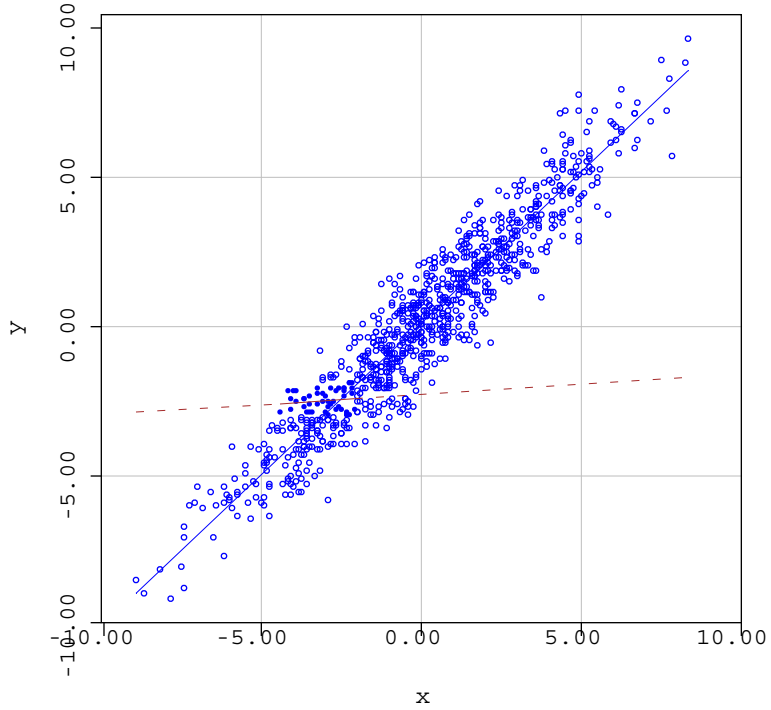
EJERCICIO 4.2.4 En el informe PISA, ¿como interpretarías o explicarías que el índice de status socieconómico y cultura y puntuación en Matemáticas tiene una forma curvilínea?

4.3.Qué podemos ver con un diagrama de dispersión

- Del mismo modo que para una sóla variable contínua hay una forma de los datos que consideramos más típica (la que se asemeja a la distribución normal) y otras que se desvian de esa forma, cuando tenemos dos variables también tenemos un ejemplo más típico y otras que se desvian de esa forma.
- Las cosas que podemos ver son:
  - Si las relaciones son positivas o negativas
  - Relaciones fuertes o débiles
  - Si las relaciones son rectas o no
  - Si hay concentraciones de datos en lugares que no son los comunes
  - Si hay valores llamativos

4.4.El ejemplo más prototípico

- Este es un ejemplo muy perfecto de un diagrama de dispersión entre dos variables (es inventado)



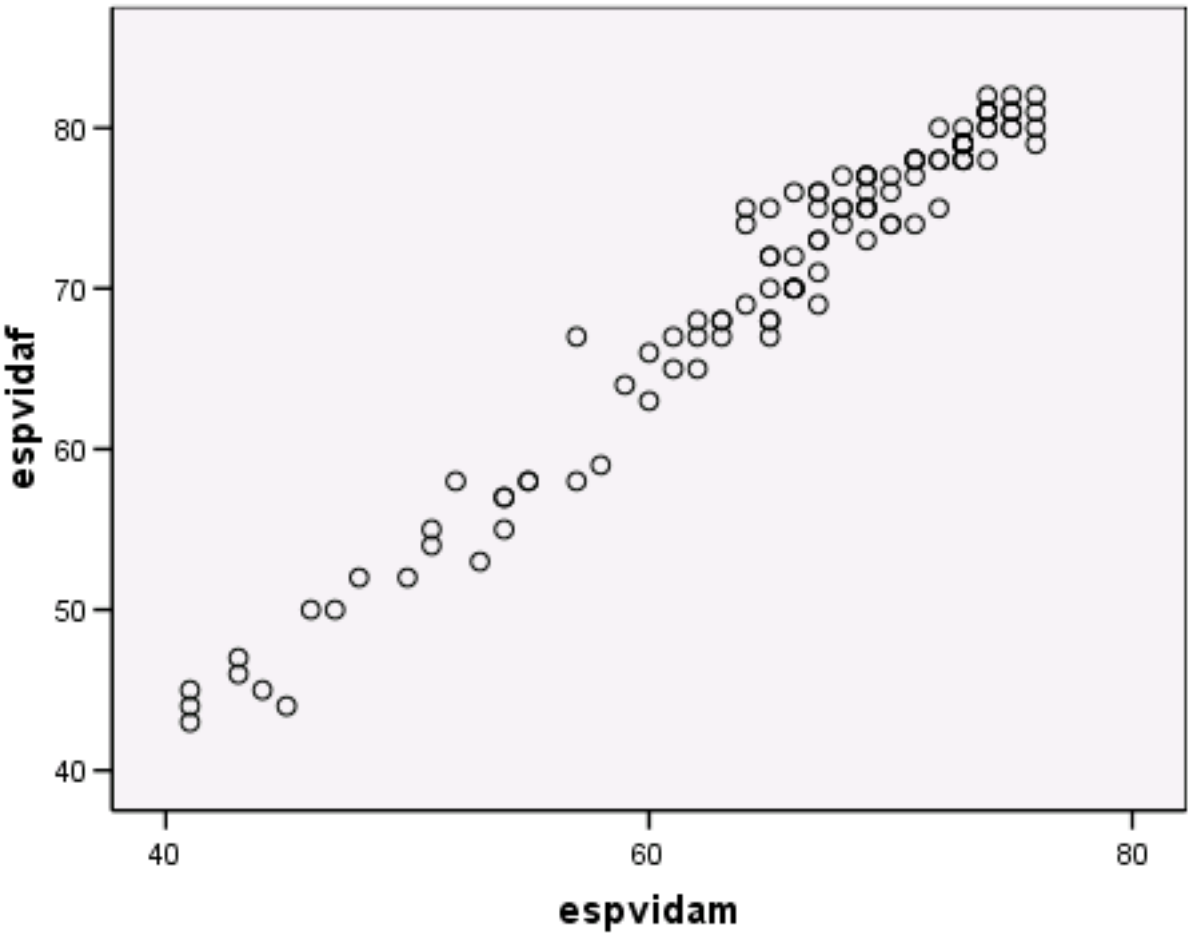
  - (La línea de puntos no hacerle caso)

- Los puntos forman como una especie de tubo.
- Los lados están más dispersos que el centro (donde hay más concentración de puntos)
- La relación es como una línea recta y la relación es positiva (cuanto más x más y)

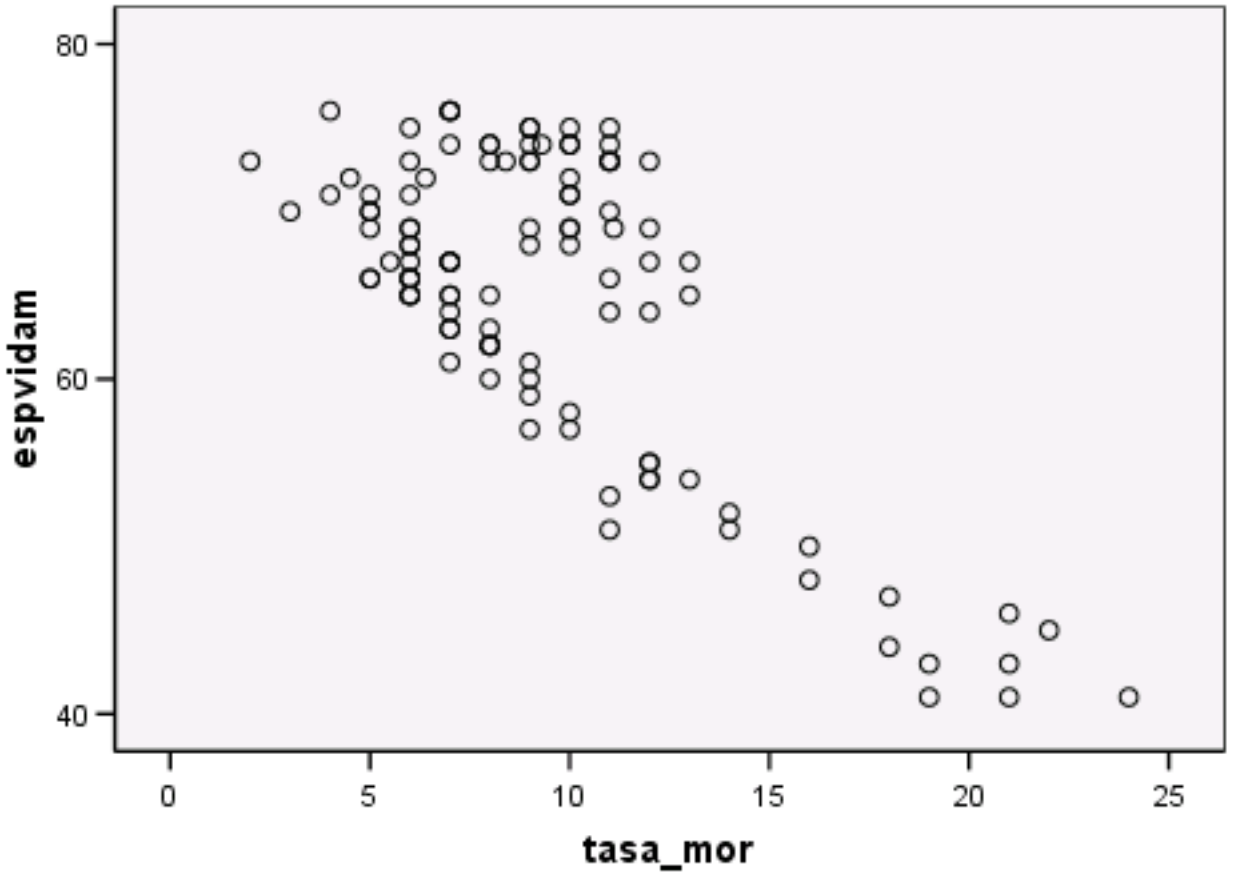
4.5.Relaciones positivas o negativas

- En una relación positiva, cuanto más de una variable, más de la otra.

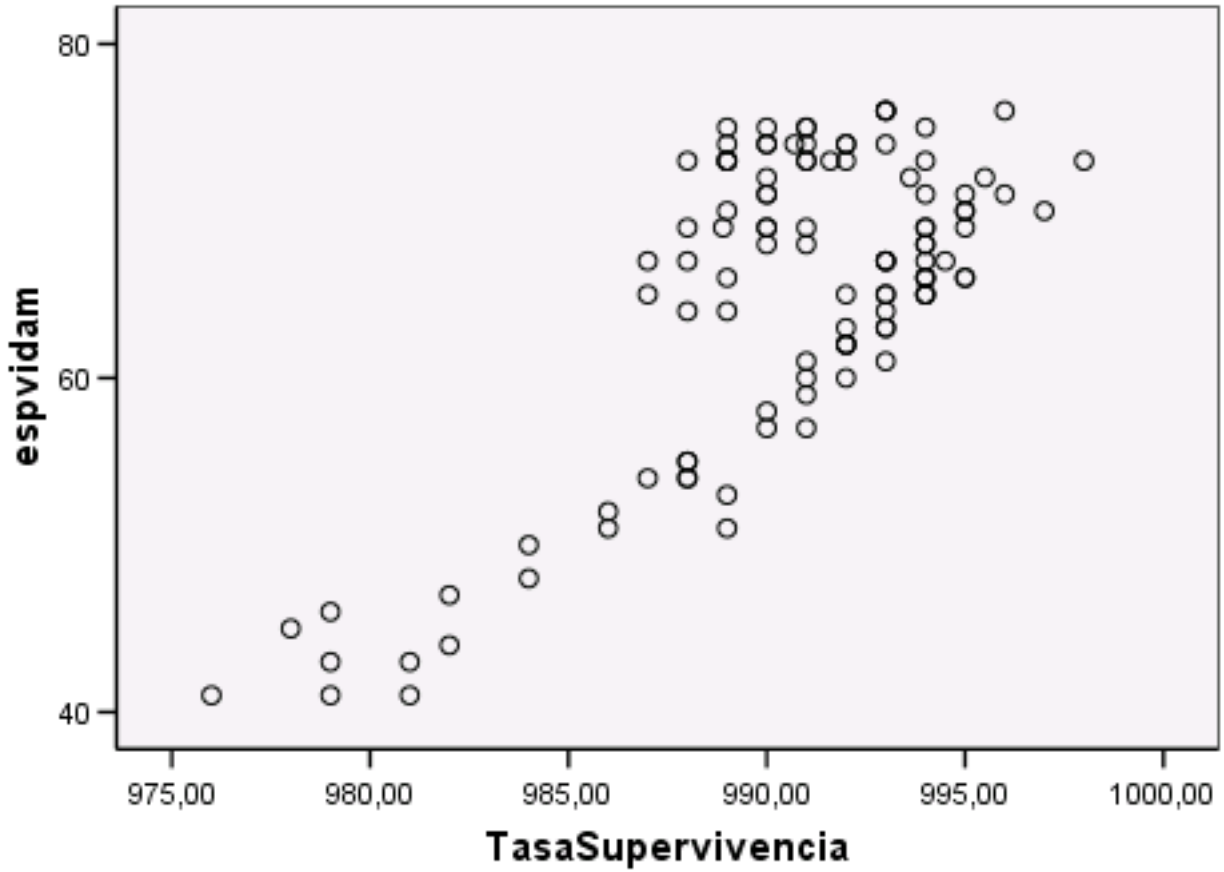
– Por ejemplo, en esperanza de vida masculina y femenina.



– En cambio, a veces la relación es negativa, como la tasa de mortalidad y la esperanza de vida masculina



- Hay que tener en cuenta que positivo o negativo es algo arbitrario, que depende de la manera en que decidamos medir las variables.Por ejemplo, si hacemos tasa de supervivencia=1000-tasa de mortalidad

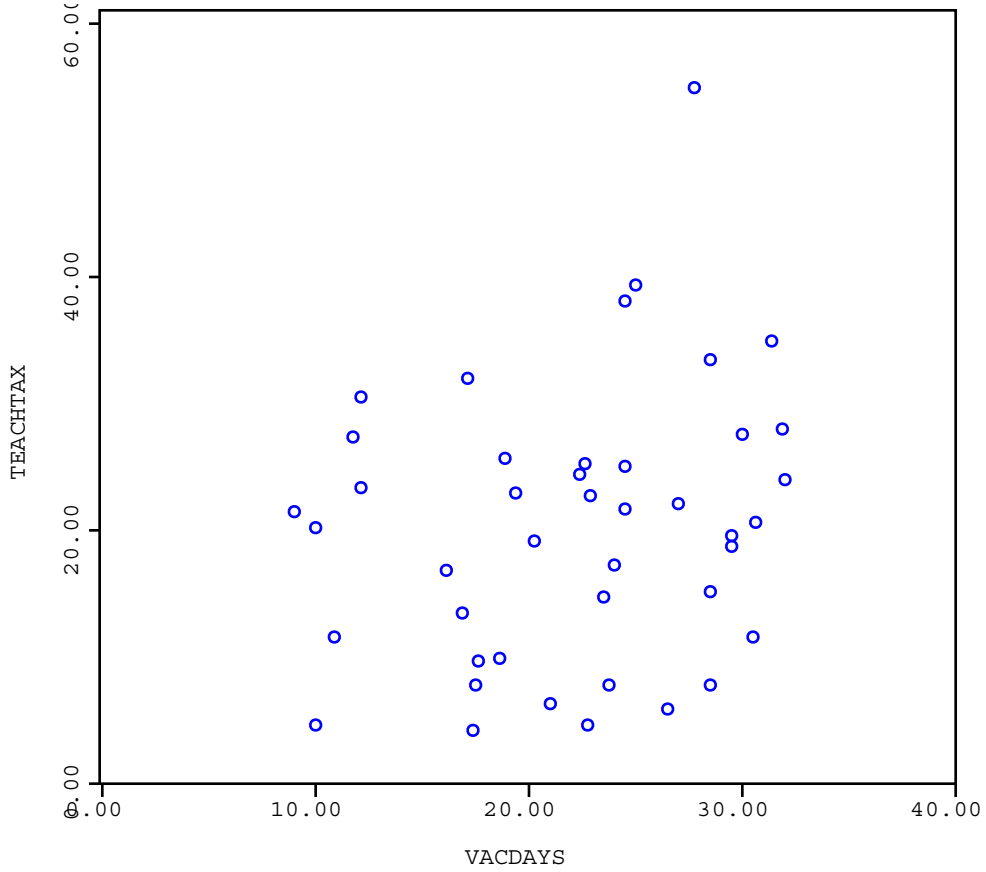


  - El gráfico es como una imágen de espejo del anterior pero el significado es el mismo, naturalmente.
- Que la relación sea positiva o negativa es una cuestión un tanto arbitraria.
  - Siempre es posible invertir una de las variables y hacer que la relación sea la inversa
  - Este tipo de inversiones a veces es desable para evitar confusiones en la interpretación

4.6.La fuerza de la relación

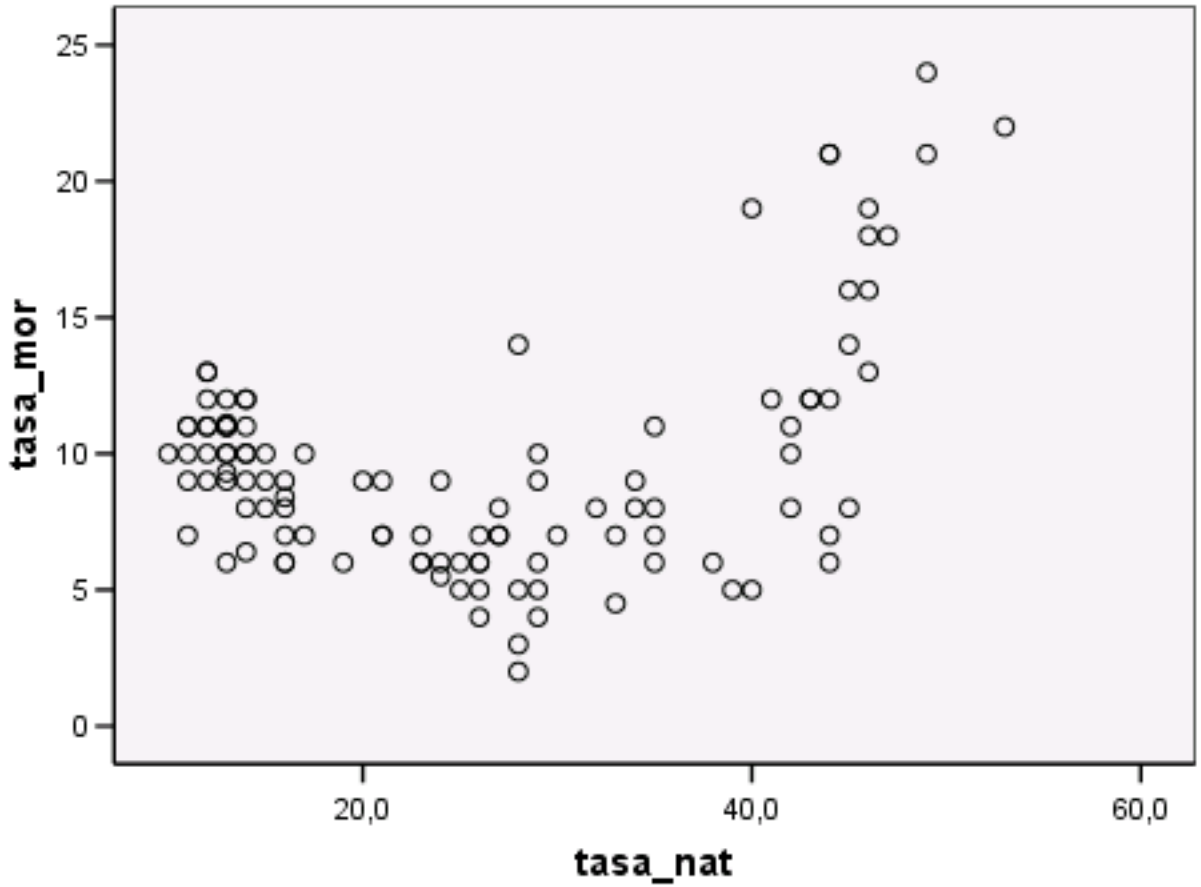
- Cuanto más estirada está la nube de datos, la relación es más fuerte.
  - Si la nube de datos parece un balón, entonces la relación es cercana a cero.

– Por ejemplo, en los datos de BigMac vemos que los dias de vacaciones no tienen relación con los sueldos (de los maestros):



4.7.Relaciones curvilineas

- Veamos la relación entre tasa natalidad y mortalidad

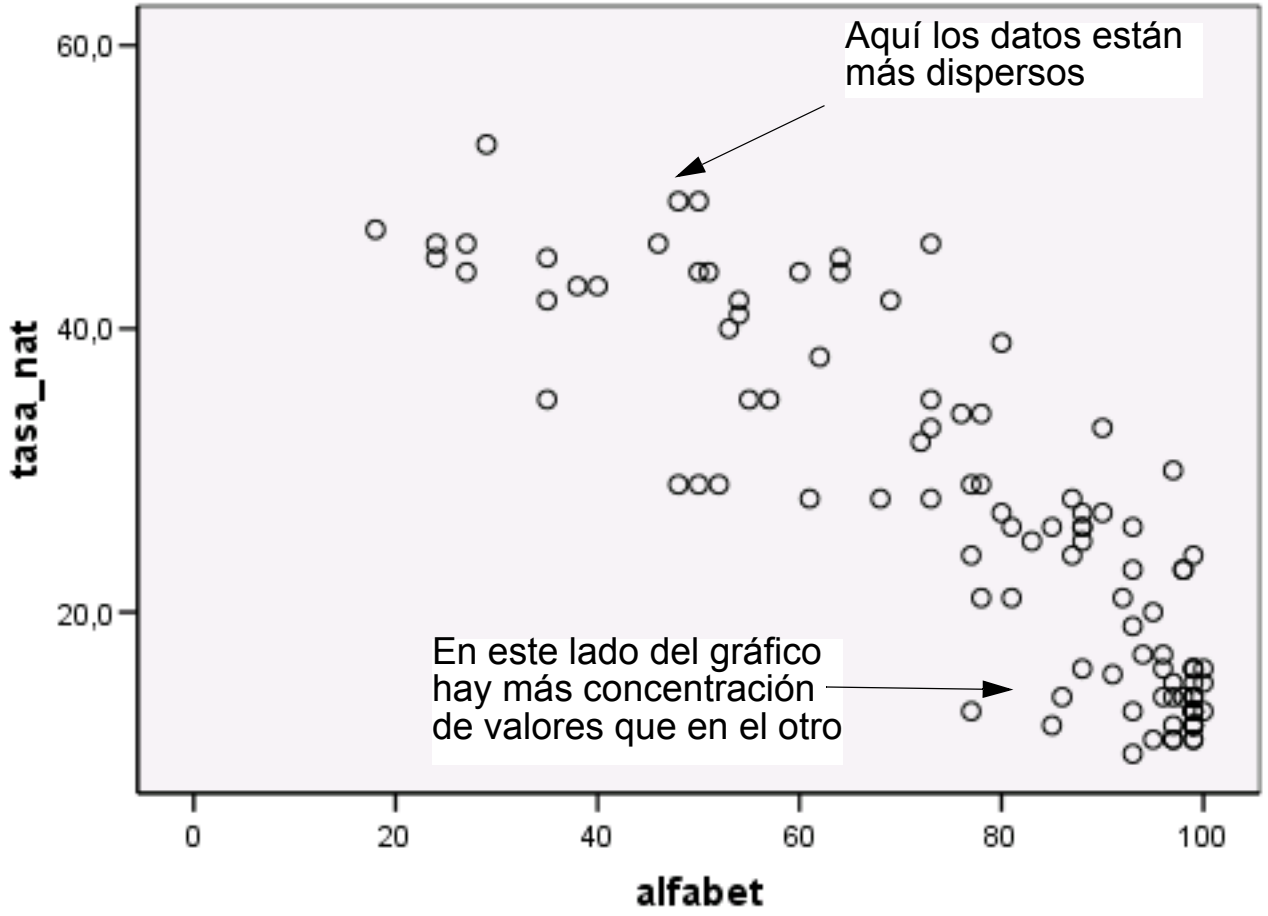


- A medida que la tasa de natalidad es mayor cuando los valores son bajos, la tasa de mortalidad desciende.
- Hay un punto en que la tasa de natalidad no parece estar relacionada con la tasa de mortalidad
- Cuando los valores de tasa de natalidad son altos, entonces la tasa de mortalidad aumenta bastante

4.8.Concentraciones en lugares inesperados

- En la versión inicial del diagrama de dispersión dijimos que la forma más idealizada es que los datos tuvieran como una forma de tubo.
  - En la parte central de los datos hay más concentración de valores
  - En los extremos del tubo hay menos concentración
- En ocasiones no obstante la concentración se puede dar en lugares que no siguen esa forma idealizada.

- Por ejemplo, si ponemos la alfabetización y la tasa de natalidad de los países tenemos:



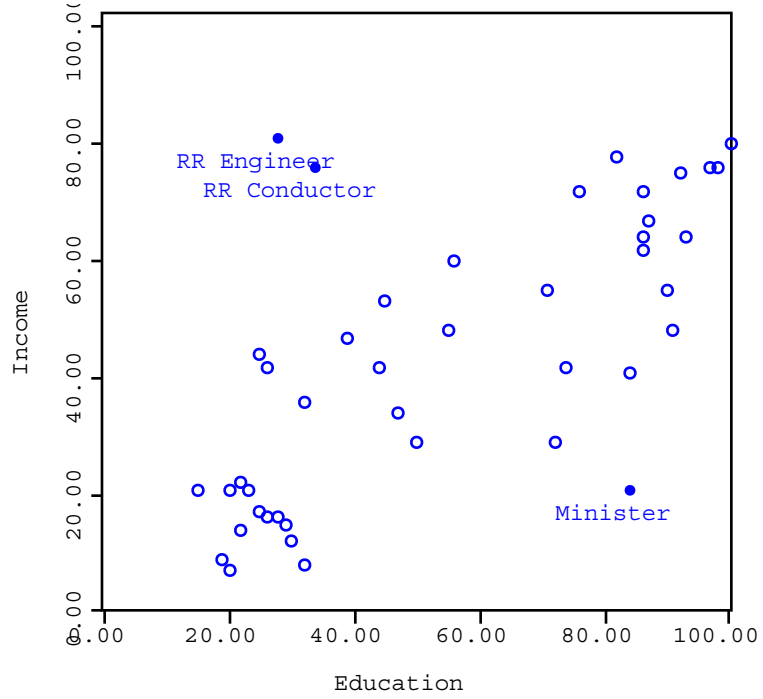
4.9.Valores llamativos o destacados

- Cuando miramos a un diagrama de dispersión a menudo podemos ver una tendencia y también puntos que se desvian mucho de esta tendencia. Esos valores llamativos son importantes ya que pueden tener información especial.

Ejemplo

En en estudio se analizaron una serie de datos acerca de profesiones (los datos son del año 1950). Tenemos nivel de ingresos (medido como el porcentaje de personas que cobraban más de 3500 dolares), la educación necesaria para alcanzarlo (medida como el porcentaje de gente que tiene esa profesión y pasó del instituto) , y el nivel de prestigio (medido como el porcentaje de gente que valoró esa profesión como excelente o muy buena) de una serie de profesiones. Un objetivo de este estudio sería ver como la educación influ-

yen en el nivel de ingresos



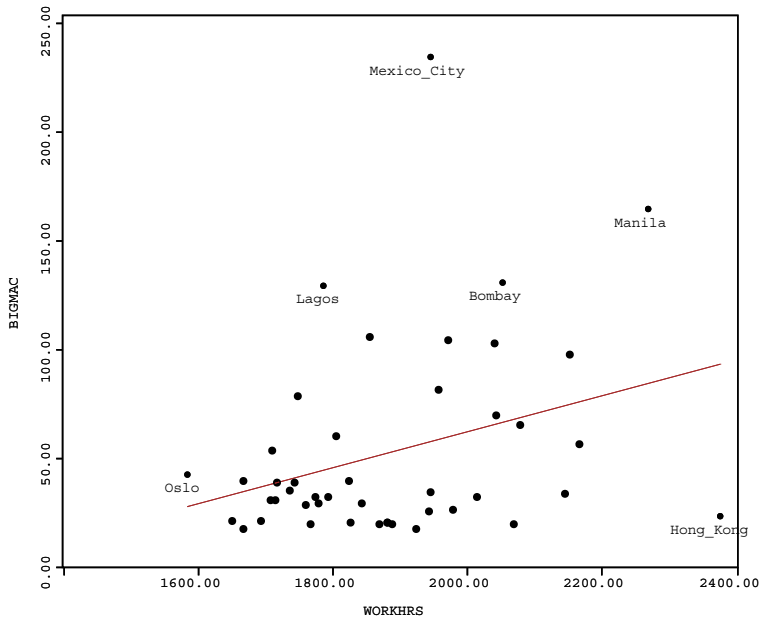
– Vemos que en general la relación es positiva pero que hay tres casos en que los puntos están un poco más alejados.

- Esos puntos corresponden a sujetos con alto nivel de ingresos para su educación, o bajo nivel de ingresos para su educación

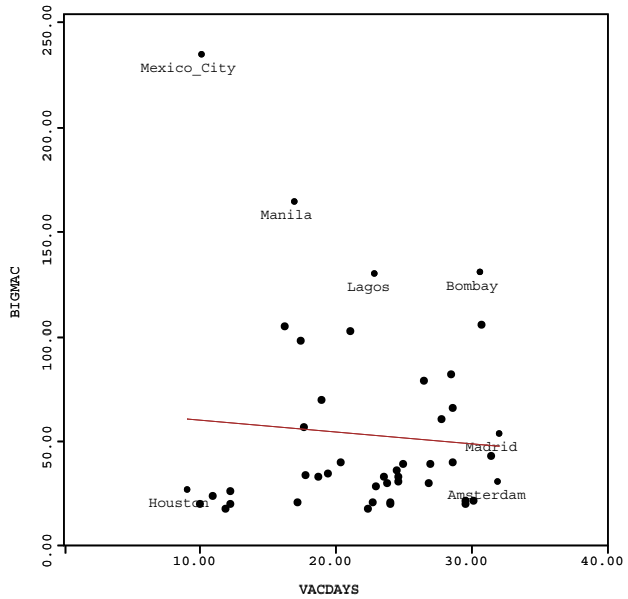
ACTIVIDADES

EJERCICIO 4.9.1 En secciones anteriores vimos el ejemplo de los datos acerca del precio de las hamburguesas. Poniendo ese precio en relación con otras variables

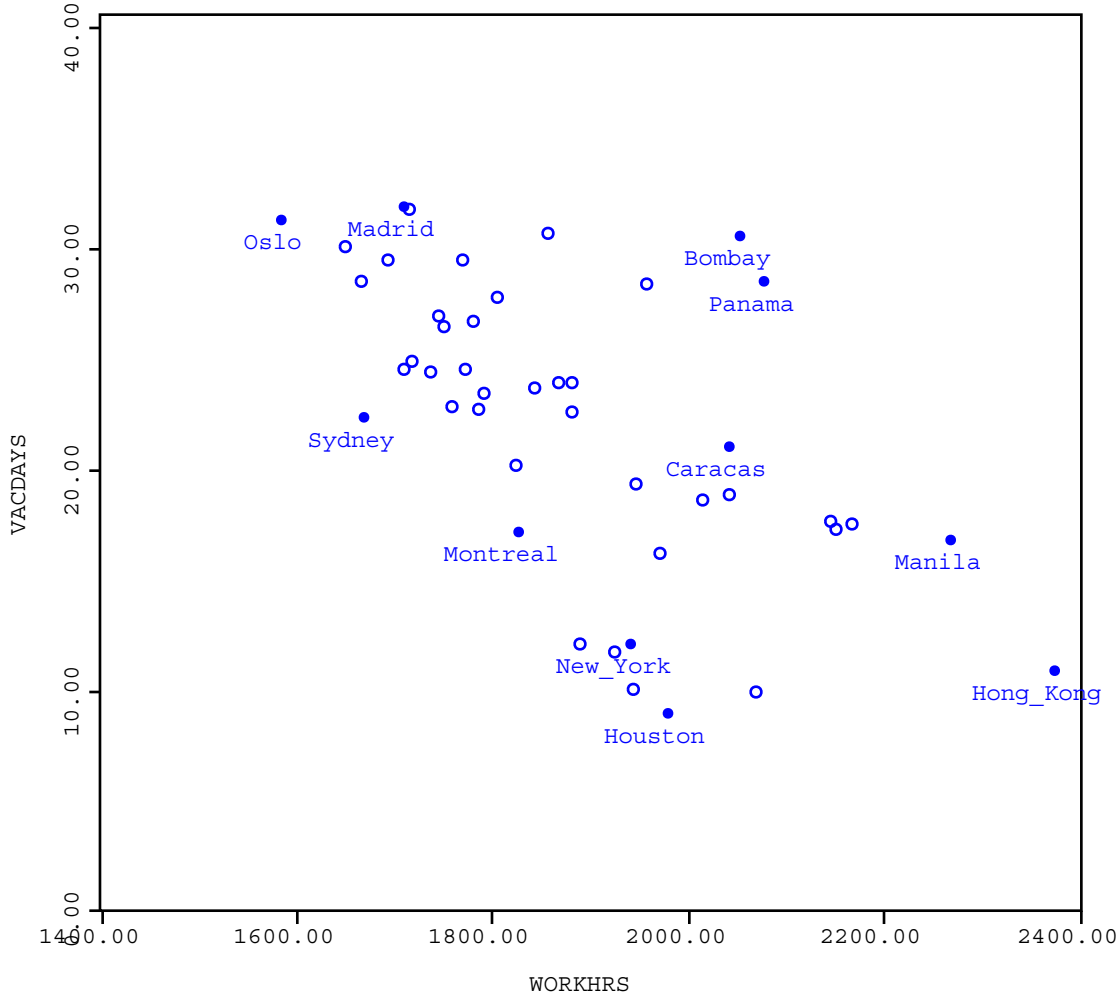
a) Relación entre horas de trabajo promedio hechas al año y precio de la hamburguesa



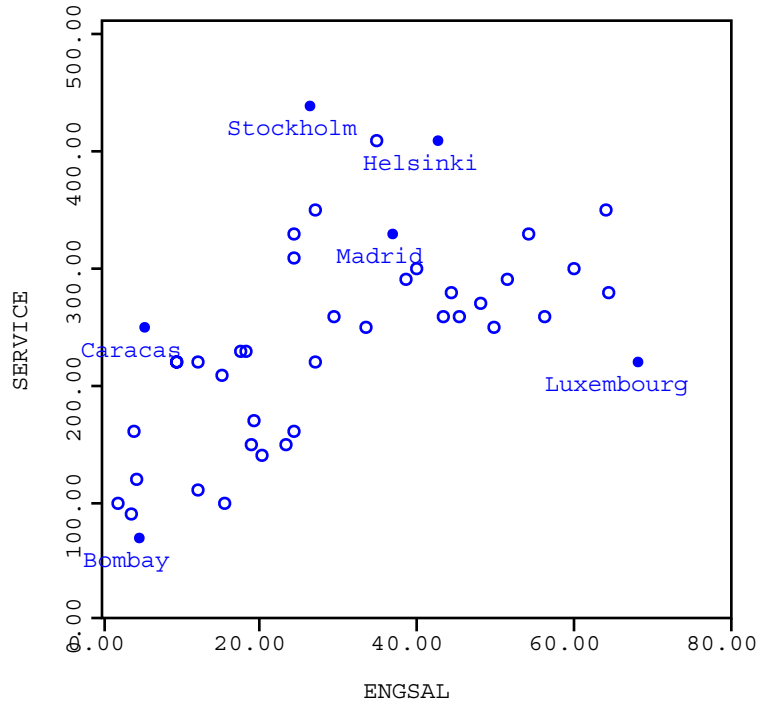
b) Relación entre días de vacaciones y precio de la hamburguesa



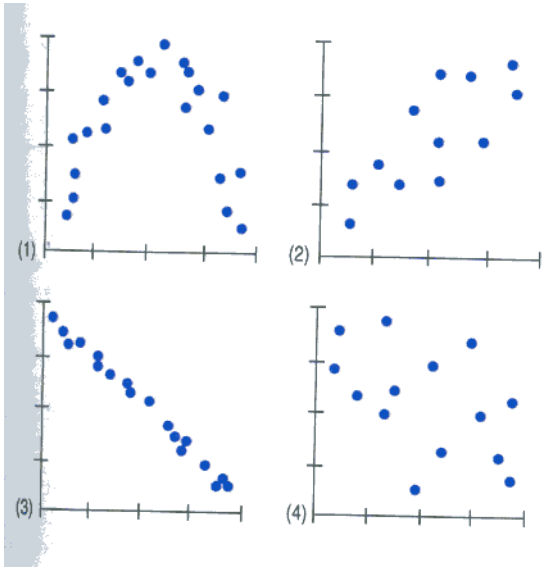
c) Relación entre días de vacaciones promedio al año y horas trabajadas promedio al año



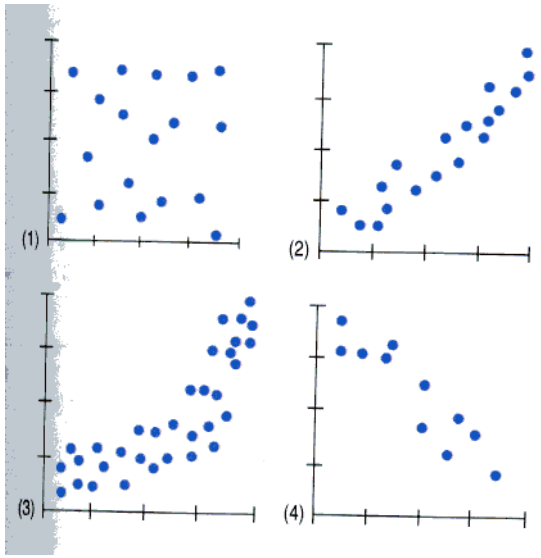
d) Relación entre sueldo (de un ingeniero) y coste de una serie de servicios (es decir, coste de la vida)



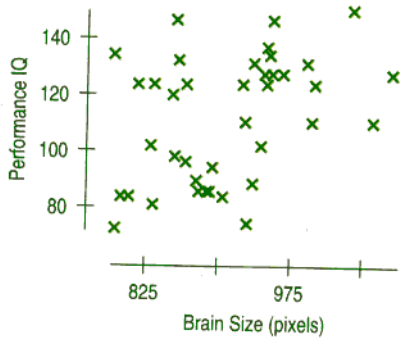
EJERCICIO 4.9.2 En estos diagramas de dispersión. ¿Cuál dirías que no hay relación entre las variables? ¿Cuál la relación es curvilínea? ¿Cuál es recta? ¿Cuál es positiva y cual negativa?



EJERCICIO 4.9.3 ¿Y de estos?



EJERCICIO 4.9.4 En este gráfico se muestra un estudio en el que se puso en relación el tamaño del cerebro de unas personas y su inteligencia medida según el Weschler ¿Dirias que hay relación entre ambas cosas mirando este gráfico?



4.10.La recta de regresión

Calculando puntuaciones predichas

- La recta de regresión nos da una idea de la relación teórica entre dos variables.
- Cuando hablamos de regresión, hay una variable explicada o predicha, y una variable explicativa o predictora (también, dependiente o independiente).
- La variable predicha se pone en el eje de las y, la variable predictora en el eje de las x.
- En el ejemplo de la Figura 1 sobre la relación entre puntuación en Matemáticas y la inversión pública, la formula no nos la dan así que usaremos otro ejemplo.

- Ejemplo, 18 esquiadores de campo a través hacen un recorrido. A esos esquiadores se les mide la concentración de CPK en sangre (la cantidad de enzima CPK en sangre es una medida de stress muscular). Los datos son los siguientes.

DataType:BiVar	Age	CPK
Size:18 X 2	Numeric	Numeric
Obs1	19.00	520.00
Obs2	21.00	300.00
Obs3	24.00	480.00
Obs4	24.00	1040.00
Obs5	24.00	1360.00
Obs6	25.00	580.00
Obs7	32.00	440.00
Obs8	33.00	180.00
Obs9	35.00	490.00
Obs10	37.00	520.00
Obs11	37.00	380.00
Obs12	44.00	640.00
Obs13	50.00	360.00
Obs14	51.00	240.00
Obs15	52.00	420.00
Obs16	55.00	280.00
Obs17	57.00	400.00
Obs18	62.00	260.00

- CPK es la variable predicha, AGE es la variable predictora
- La formula en este caso es la siguiente:

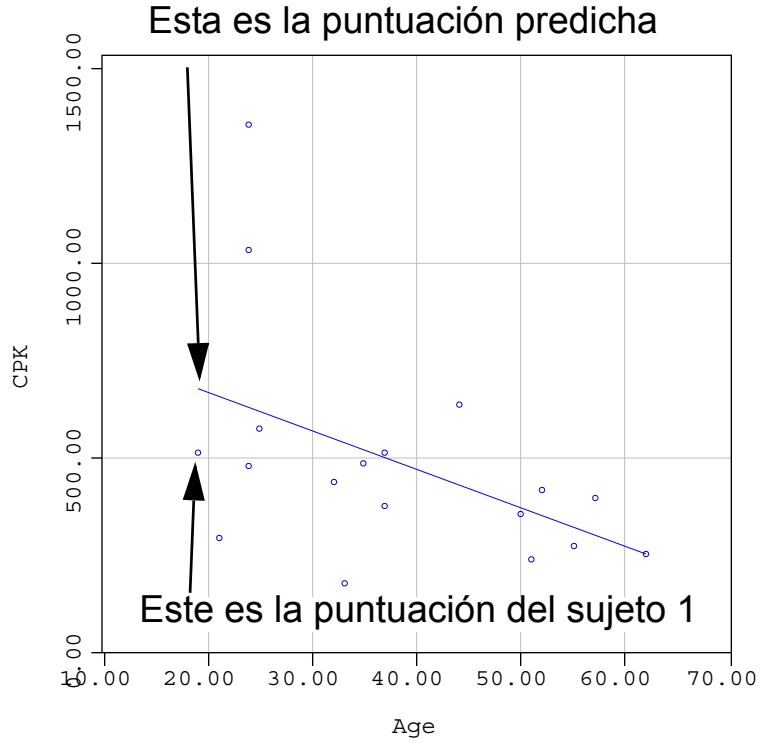
$$CPK = 867 - 9.85 \times AGE$$ **Ecuación (2)**

  - A partir de esa fórmula podemos calcular las puntuaciones predichas o esperadas para el CPK de los esquiadores.

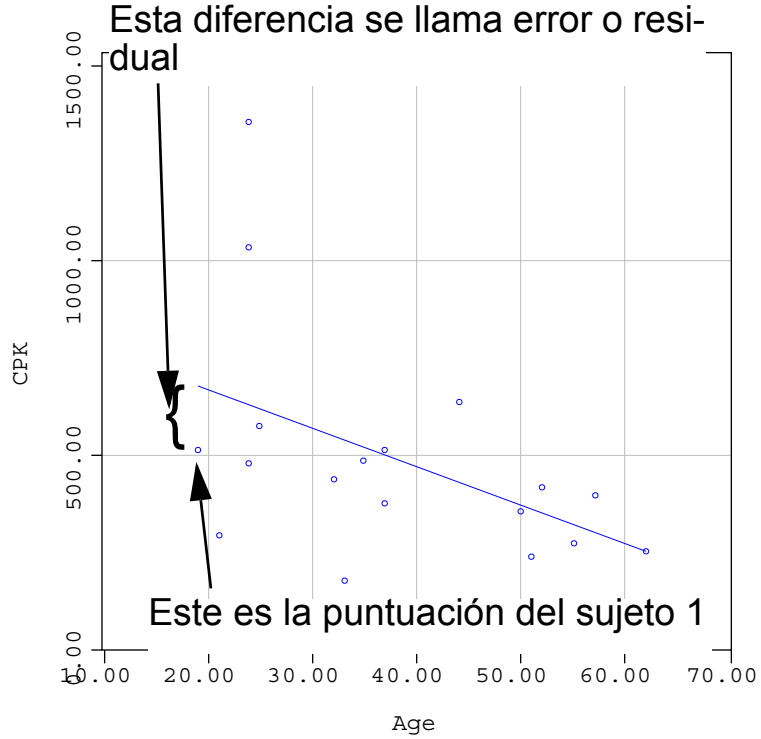
– Cada puntuación expresa un valor teórico o ideal que asignamos a todos los sujetos que tengan el mismo valor en la variable predictora. Por ejemplo, el primer esquiador tenía una edad de 19. El valor predicho para este esquiador es de:

$$679.85 = 867 - 9.85 \times 19$$ **Ecuación (3)**

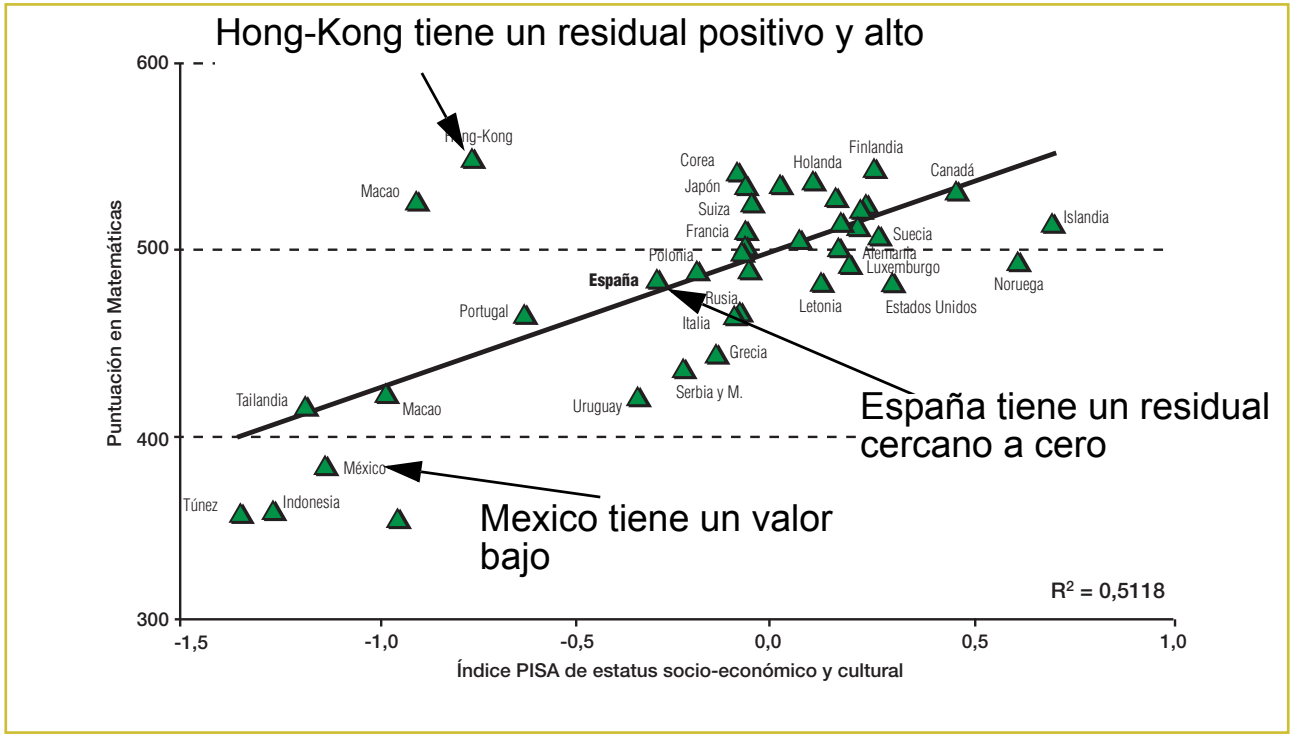
– El gráfico a continuación muestra la línea de regresión y el valor del primer sujeto.



- Las puntuaciones predichas y las observadas no coinciden. Siempre hay una cierta cantidad de error.



- El concepto de error o residual es de mucha importancia. En el informe PISA lo utilizamos para valorar si un país está funcionando por encima o por debajo de sus posibilidades o expectativas.



Fuente: *PISA*

Figura 2: Diagrama de Dispersión dePuntuación en Matemáticas versus estatus socioeconómico

- Para calcular los residuales simplemente restamos la puntuación observada de la puntuación predicha. Los símbolos que se suelen utilizar son:

$$e_i = y_i - \hat{y}_i \qquad \textbf{Ecuación (4)}$$

Esto indica el residual o error

Esta es la puntuación predicha

Esta es la puntuación observada

- Las puntuaciones observadas, predichas y residuales para los datos de los esquiadores son las siguientes:

Observa- Predi- Residua-

DataType:MulVar	Response	Fit Value	Residuals
Size:18 X 8	Numeric	Numeric	Numeric
Obs1	520.00	679.90	-159.90
Obs2	300.00	660.20	-360.20
Obs3	480.00	630.66	-150.66
Obs4	1040.00	630.66	409.34
Obs5	1360.00	630.66	729.34
Obs6	580.00	620.81	-40.81
Obs7	440.00	551.88	-111.88
Obs8	180.00	542.03	-362.03
Obs9	490.00	522.34	-32.34
Obs10	520.00	502.64	17.36
Obs11	380.00	502.64	-122.64
Obs12	640.00	433.71	206.29
Obs13	360.00	374.62	-14.62
Obs14	240.00	364.78	-124.78
Obs15	420.00	354.93	65.07
Obs16	280.00	325.39	-45.39
Obs17	400.00	305.69	94.31
Obs18	260.00	256.45	3.55

Tabla 12: Puntuaciones observadas, predichas y residuales para los datos de esquiadores

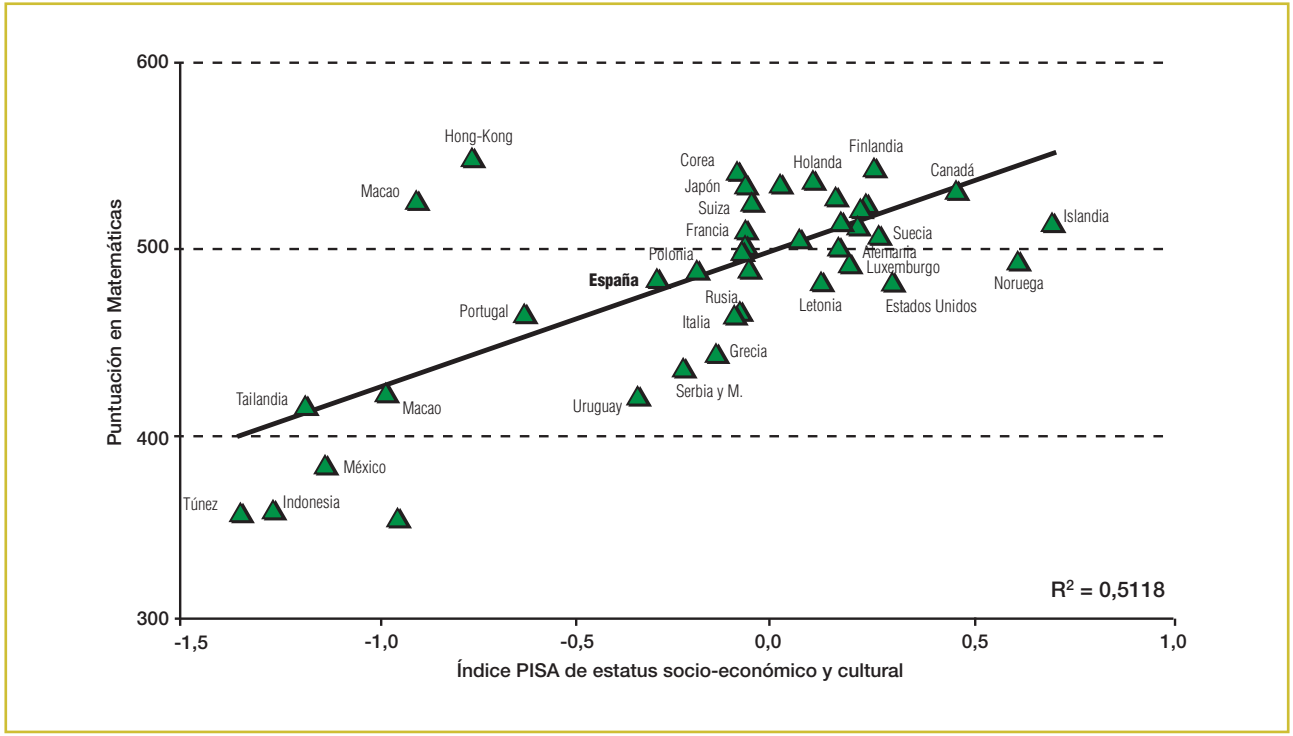
- Tener en cuenta las siguientes relaciones. Las tres fórmulas son la misma después de un poco de manipulación

$$
\begin{aligned}
e &= y - \hat{y} \\
y &= \hat{y} + e \\
\hat{y} &= y - e
\end{aligned}
$$

ACTIVIDADES

EJERCICIO 4.10.1 Comprueba que las puntuaciones residuales de la Figura 12 están bien calculadas a partir de las otras puntuaciones. Utiliza la Ecuación 4.

EJERCICIO 4.10.2 En el gráfico de la Figura 3 indica aproximadamente cual es la puntuación observada, la predicha y la residual para Islandia. Indica lo mismo para Macao.



Fuente: *PISA*

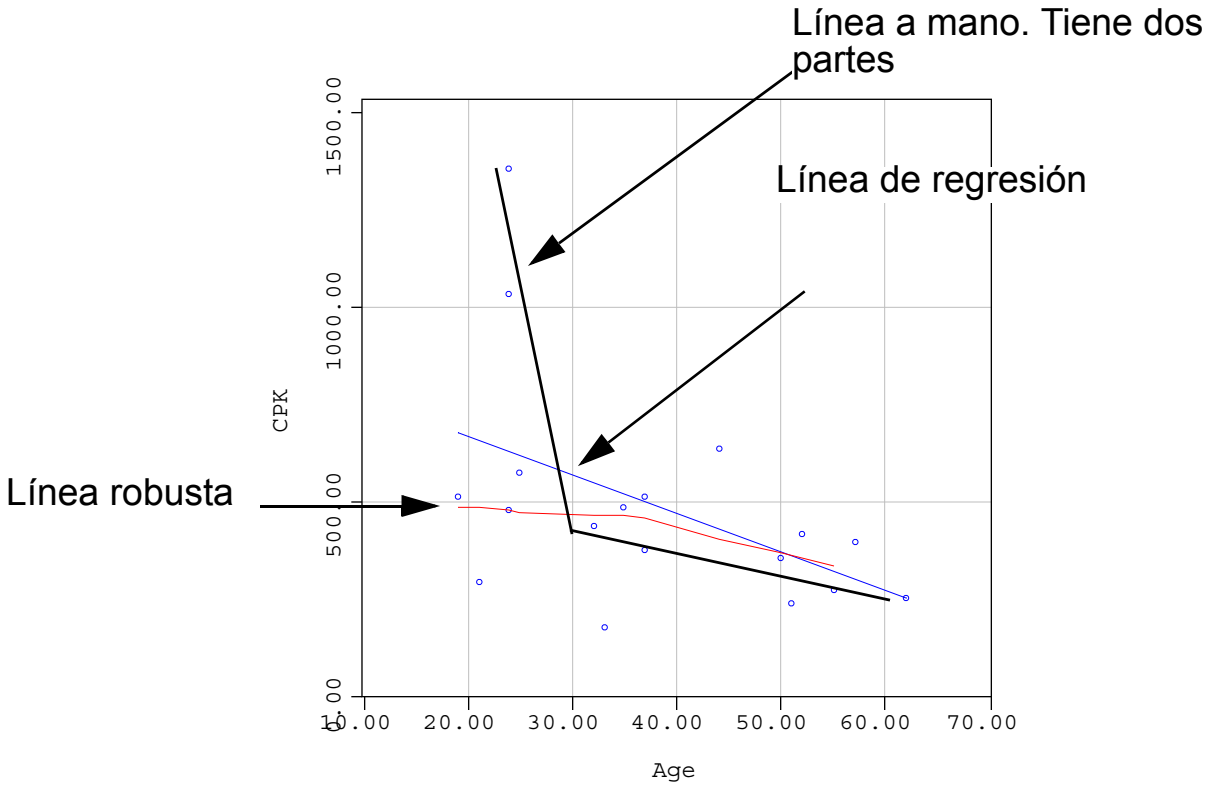
Figura 3: Diagrama de Dispersión dePuntuación en Matemáticas versus inversión Pública

4.11.Como calcular rectas (1)

Métodos para ajustar líneas

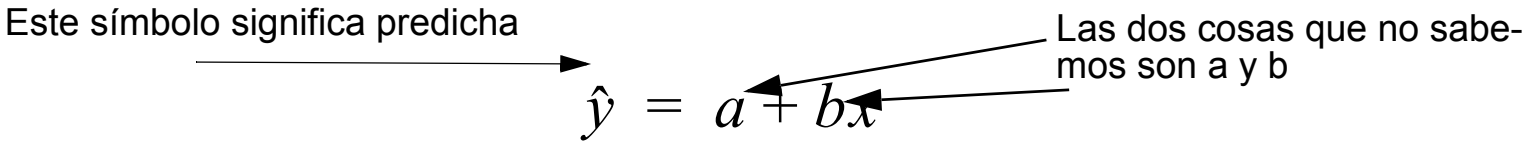
- Hay diversas maneras de ajustar líneas de predicción:
  - Se pueden ajustar a mano
  - Se pueden calcular rectas de regresión
  - Se pueden utilizar métodos robustos o más flexibles

- Aquí se muestran varias posibilidades



- Cada una de estas líneas tiene sus propiedades y sus méritos. Nosotros sólo veremos las de la regresión.

- ***La línea de regresión*** tiene las siguientes propiedades:
  - Es recta
  - Siendo recta, es la línea que da una suma al cuadrado de residuales menor.
- Hay varias formulas para calcular la recta de regresión. De entre ellas, he seleccionado la siguiente:
  - Recordar, queremos calcular una recta que tiene la siguiente forma:



– b se denomina la pendiente de la recta y se puede calcular con la siguiente fórmula.

$$b = \frac{rS_y}{S_x}$$

r=correlación
S_y = desviación típica de y
S_x = desviación típica de x

Ecuación (5)

– Una vez se conoce b es fácil calcular a. Esto se hace mediante

$$a = \bar{y} - b\bar{x}$$

Ecuación (6)

– En los cálculos anteriores falta como calcular r. Eso lo veremos en el siguiente apartado.

4.12.El coeficiente de correlación

La fuerza de la relación

- Hay diferentes fórmulas para calcular el coeficiente de correlación, todas ellas equivalentes.
- El método más sencillo consiste en:
  - Pasar las puntuaciones de las variables a puntuaciones z (esto se hace restando la media de la variable y dividiendo por la desviación típica)

$$z_x = \frac{x_i - \bar{x}}{s_x}$$

Ecuación (7)

– Aplicar la siguiente fórmula (es decir, multiplicar cada una de las puntuaciones z para una variable por la puntuación z correspondiente de la otra variable y dividir por el número de casos)

$$r = \frac{\sum z_x z_y}{n}$$

Ecuación (8)

Ejemplo

EJEMPLO DE CALCULO: Se llevó a cabo un estudio de efecto del carril-bici sobre conductores y ciclistas. Las variables son: ESPACIO DE VIAJE entre el carril-bici y la línea central de la carretera, y SEPARACIÓN es la distancia entre el ciclista y un coche que pasa.

Ejemplo

Aquí están los datos

Tabla 13: Datos para Espacio de viaje y separación

Espacio de viaje	Separa-ción
12.8	5.5
12.9	6.2
12.9	6.3
13.6	7
14.5	7.8
14.6	8.3
15.1	7.1
17.5	10
19.5	10.8
20.8	11

Ejemplo

Calculamos la media y la desviación típica de las variables

Tabla 14: Medias y desviaciónes típicas de las variables

	$\bar{x}$	S
Espacio de viaje	**15.42**	**2.88**
Separación	**8**	**1.98**

Ejemplo

Calculamos las puntuaciones típicas

Tabla 15: Puntuaciones z para Espacio de viaje y separación

Espacio de viaje	Separa-ción
-0.91	-1.26
-0.88	-0.91
-0.88	-0.86
-0.63	-0.51
-0.32	-0.1
-0.29	0.15
-0.11	-0.46
0.72	1.01
1.42	1.42
1.87	1.52

Ejemplo

Multiplicamos los dos valores de cada fila

Tabla 16: Multiplicando los valores

Multiplica-ción
1.15
0.80
0.76
0.32
0.03
-0.04
0.05
0.73
2.02
2.84

Ejemplo

Sumamos los valores de la columna anterior y dividimos por el número de casos y nos da la correlación.

Tabla 17: Correlación

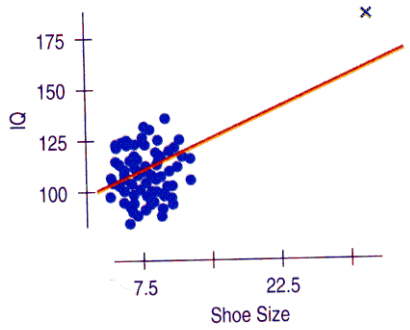
Correla-ción
0.96

4.13.Algunas propiedades de los coeficientes de correlación

- Los coeficientes de correlación pueden estar entre -1 y 1
  - Si el valor es positivo la relación es positiva (a más de una cosa, más de la otra)
  - Si el valor es negativo la relación es inversa o negativa (a más de una cosa menos de la otra)
  - Recordar lo que vimos sobre invertir relaciones en la Sección 4.5.
- Una relación de 1 o -1 indica una relación perfecta. Todos los puntos caen en la línea

- Valorar si una correlación es alta o baja es algo relativo al resto de las correlaciones o a otros factores.
  - Si todas las correlaciones que obtenemos están entre 0 y 0.1, una correlación de 0.4 nos puede parecer muy alta
  - Si hay muchos factores que pueden oscurecer la relación pero aun así obtenemos unas correlaciones moderadas, entonces podemos darle mucha importancia al resultado obtenido
  - En resumen, valorar una correlación en el vacio, sin hacer referencia a situaciones concretas puede ser sin sentido.

- Hay que mirar los gráficos para evaluar una correlación. Por ejemplo, la correlación para estos datos saldría bastante alta.



4.14.Como calcular rectas (2)

Ahora sí

- En la sección Como calcular rectas (1). vimos una fórmulas para calcular la recta de regresión que utilizaba el coeficiente de correlación. Estas fórmulas eran la Ecuación 5 y la Ecuación 6.
- Tener en cuenta que, a diferencia de la correlación, tenemos que distinguir entre variable predictora y variable predicha. En nuestro ejemplo, el espacio de viaje es la predictora, y la separación es la predicha.

- Aplicado al ejemplo, el resultado es:

$$b = \frac{0.96(1.98)}{2.88} = 0.66$$

$$a = 8 - 0.66(15.42) = -2.1772$$

ACTIVIDADES

EJERCICIO 4.14.1 Para los siguientes datos, calcula la correlación del peso sobre las abdominales.

DataType:MulVar	Peso	Cintura	Abdominal
Size:19 X 3	Numeric	Numeric	Numeric
Obs1	191.00	36.00	162.00
Obs2	189.00	37.00	110.00
Obs3	193.00	38.00	101.00
Obs4	162.00	35.00	105.00
Obs5	189.00	35.00	155.00
Obs6	182.00	36.00	101.00
Obs7	211.00	38.00	101.00
Obs8	167.00	34.00	125.00
Obs9	176.00	31.00	200.00
Obs10	154.00	33.00	251.00
Obs11	169.00	34.00	120.00
Obs12	166.00	33.00	210.00
Obs13	154.00	34.00	215.00
Obs14	193.00	36.00	70.00
Obs15	202.00	37.00	210.00
Obs16	176.00	37.00	60.00
Obs17	157.00	32.00	230.00
Obs18	156.00	33.00	225.00
Obs19	138.00	33.00	110.00
New Obs			

Tabla 18: Peso, tamaño de cintura y número de abdominales de un grupo de atletas

El resultado es -0.37.

EJERCICIO 4.14.2 Calcula la correlación entre la cintura y el número de abdominales.

El resultado es -0.62.

EJERCICIO 4.14.3 Calcula la recta de regresión para predecir las abdominales a partir del peso.

El resultado es $b = -1.14$
$a = 350.15$

EJERCICIO 4.14.4 Calcula la recta de regresión para predecir las abdominales a partir de la cintura.

El resultado es

$$b = -18.18$$
$$a = 784.02$$

4.15.El ajuste de la recta

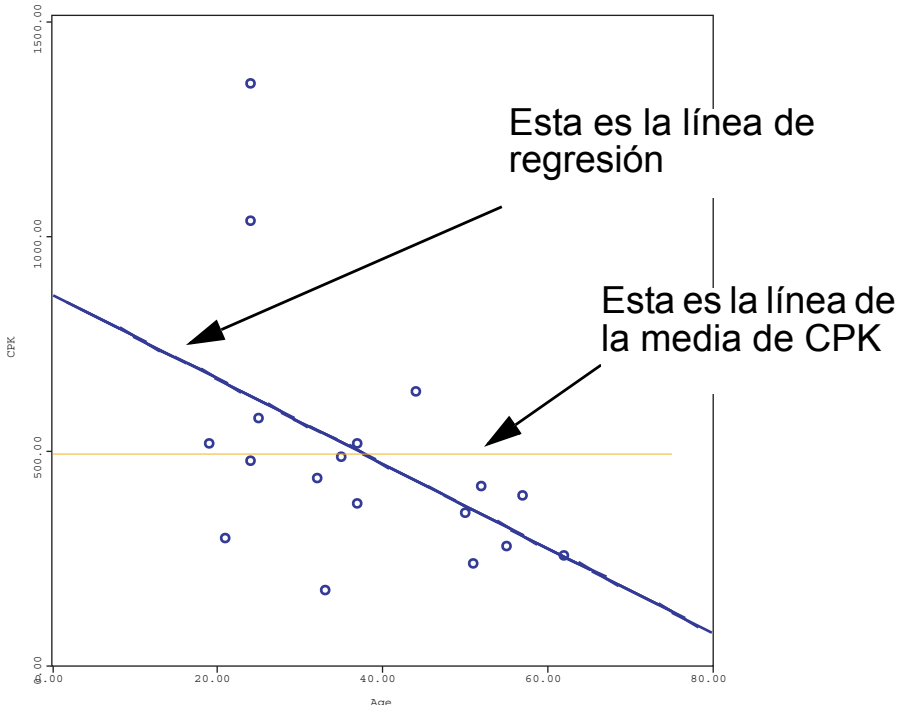
Valorando la regresión

- En la Sección 4.11. vimos que la recta de regresión es la que minimiza:

$$SCE = \sum e^2$$ **Ecuación (9)**

  - Es decir, la suma de cuadrados de los errores.
- ¿Por qué elevamos al cuadrado?
  - La suma de los residuales sin más es igual a cero. Al elevar al cuadrado los signos negativos desaparecen.

- ¿Cómo podemos valorar si SCE es mucho o es poco?
  - Volviendo al ejemplo de Edad versus CPK, tenemos lo siguiente



- La línea media es la línea recta que produce el error máximo. Si calculamos los residuales de esa línea hacemos:

$$SCT = \sum (y - \bar{y})^2$$ **Ecuación (10)**

  - Esa fórmula es igual a la de la varianza de y pero sin dividir por el número de casos.
- Sabiendo SCE y SCT podemos calcular una nueva cantidad que llamaremos suma de cuadrados explicados por la regresión (SCR).

$$SCR = SCT - SCE$$ **Ecuación (11)**

- Para valorar el tamaño de SCR calculamos la proporción (es decir, dividimos el valor más pequeño por el total). A esto lo llamamos proporción de varianza explicada y se simboliza R^2.

$$R^2 = \frac{SCR}{SCT}$$ **Ecuación (12)**

- Algunos datos sobre R^2
  - R^2 es el cuadrado de la correlación
  - R^2 va entre 0 y 1, donde uno indicaría que todos los puntos caen sobre la recta, y 0 indicaría que la recta no ayuda a mejorar la predicción en absoluto.
  - R^2 a veces se da en términos de porcentajes. Simplemente multiplicamos la proporción por 100.

– Dependiendo de la disciplina, los investigadores consideran que una R^2 es buena o mala. En encuestas, por ejemplo, una R^2 de 0.4 podría estar muy bien considerada. En experimentos físicos, un 0.9 puede ser considerado insuficiente.

ACTIVIDADES

EJERCICIO 4.15.1 Calcula el valor de R^2 para la recta de regresión que predice las abdominales a partir del peso.

EJERCICIO 4.15.2 Calcula el valor de R^2 para la recta de regresión que predice las abdominales a partir de la cintura.

EJERCICIO 4.15.3 Calcula el valor de R^2 para los datos de la Tabla 13.

Parte V
Supuestos en el cálculo de rectas de regresión

5.1.Evaluando la regresión en detalle

Calcular la proporción de varianza explicada no es suficiente

- El valor de R^2 es importante para valorar una regresión, pero también hay que tener en cuenta otras cosas. Las cosas que hay que tener en cuenta son:
  - Evaluar si la relación es realmente lineal
  - Evaluar si hay residuales de tamaño excesivo
  - Evaluar puntos influyentes
  - Considerar si hay variables subyacentes

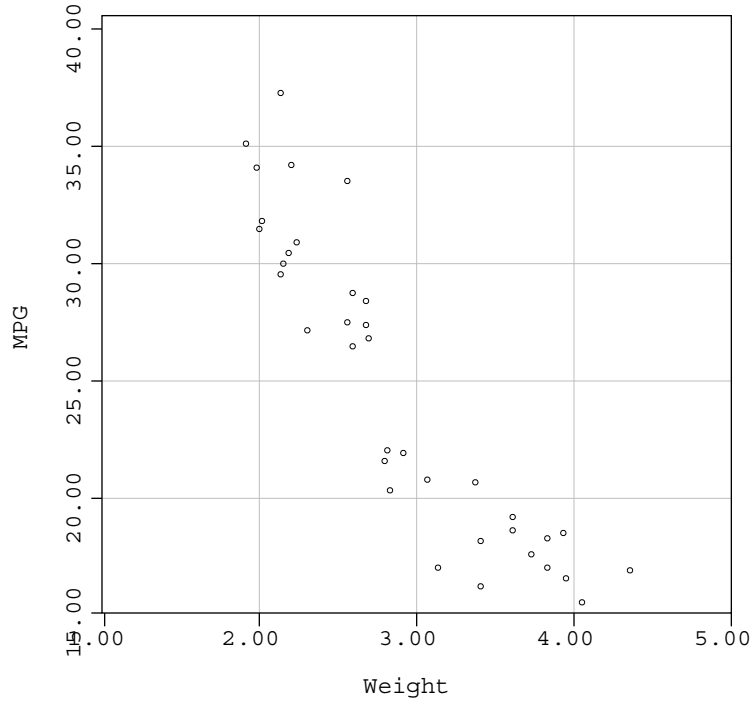
- Para evaluar lo anterior utilizaremos fundamentalmente dos herramientas.
  - El diagrama de dispersión (ya hemos visto esto antes)
  - Gráficos de los residuales: Veremos ejemplos de estos gráficos en los siguientes apartados.

5.2.Evaluar si la regresión es lineal

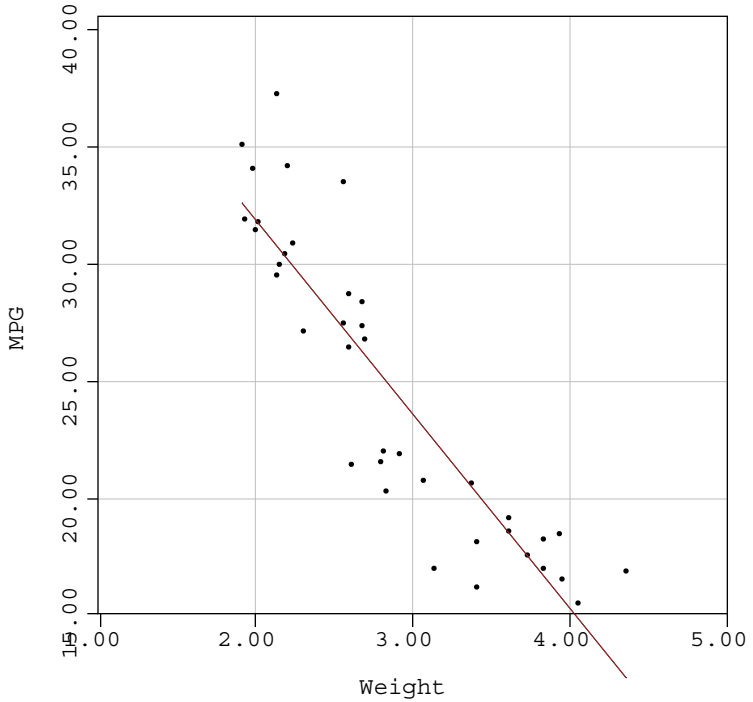
La regresión ajusta líneas rectas

- En el siguiente ejemplo examinamos 38 coches de los años 80 y vemos la relación que hay entre su peso (weight) y su eficiencia (MPG= millas por galón que es equivalente a kilómetros recorridos por litro).

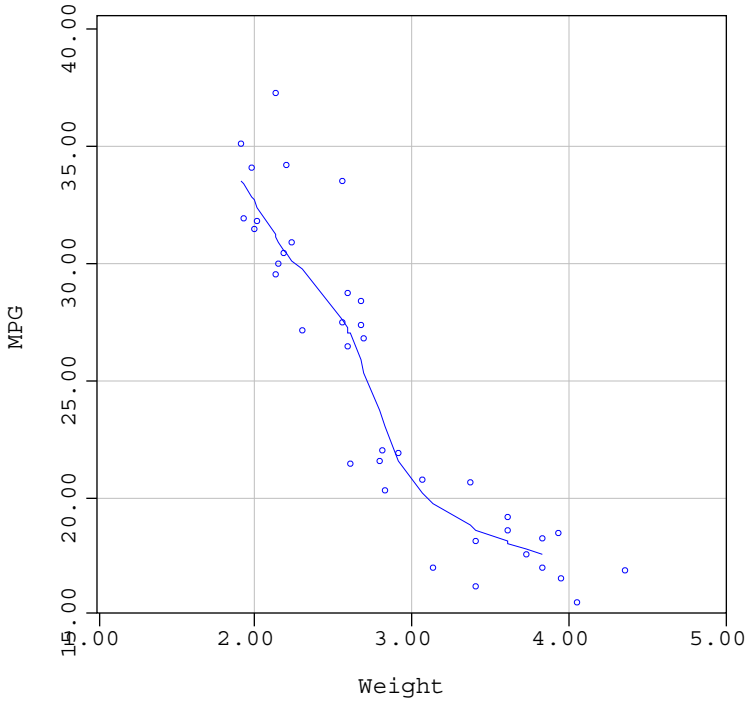
- El diagrama de dispersión sin la recta de regresión tiene este aspecto:



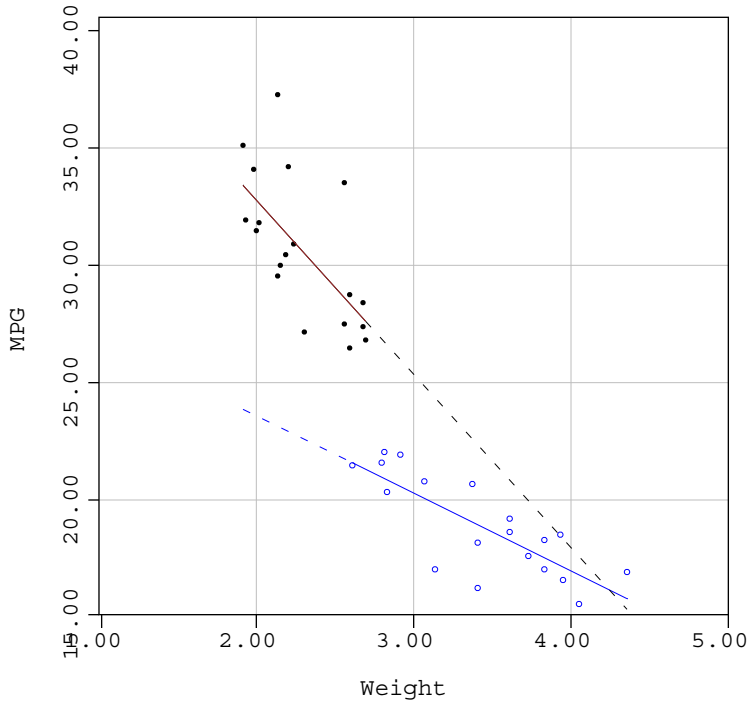
- Si ajustamos la recta de regresión veríamos esto.



- Esa línea no acaba de ajustar bien. Una línea un poco curva iría mucho mejor:



- Todavía mejor sería usar ***dos líneas rectas:***

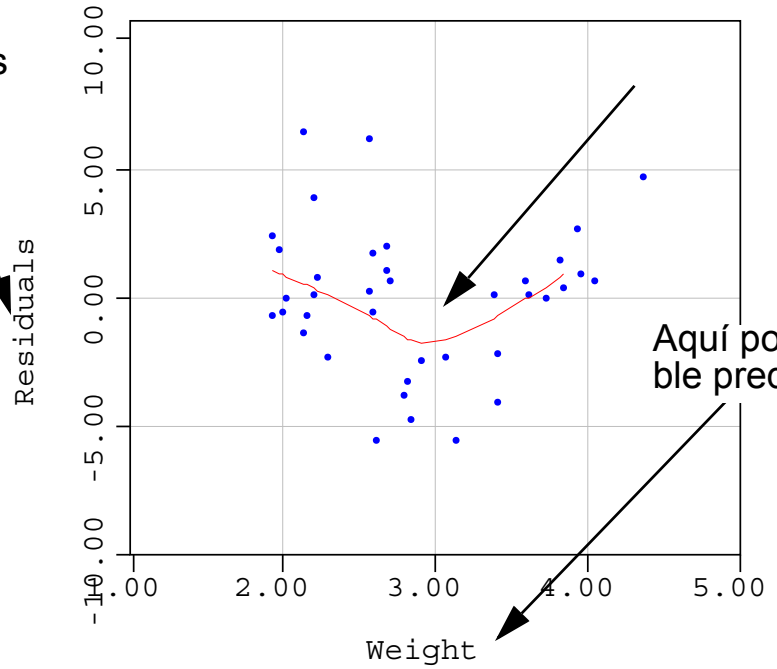


- Cuando la forma de la relación no es lineal, utilizar una recta de regresión puede no ser correcto
  - Si la relación es curvilínea, una línea recta no es una descripción adecuada de los datos
  - Si en los datos parece que hay más de un grupo, ajustar líneas por grupos puede ser más razonable

- Un gráfico para evaluar la curvilinearidad es el de la variable predictora frente a los residuales o errores

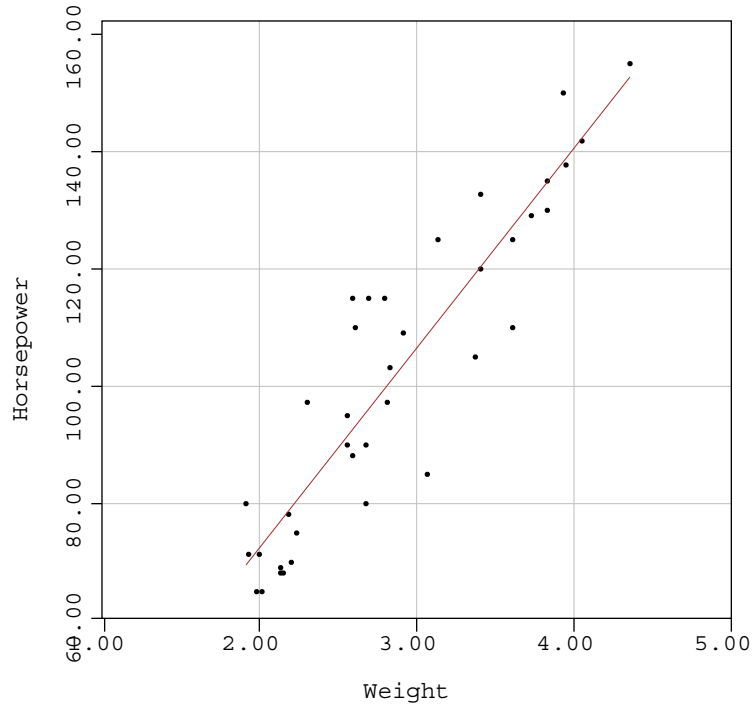
El gráfico muestra la curva muy claramente

Aquí ponemos los residuales

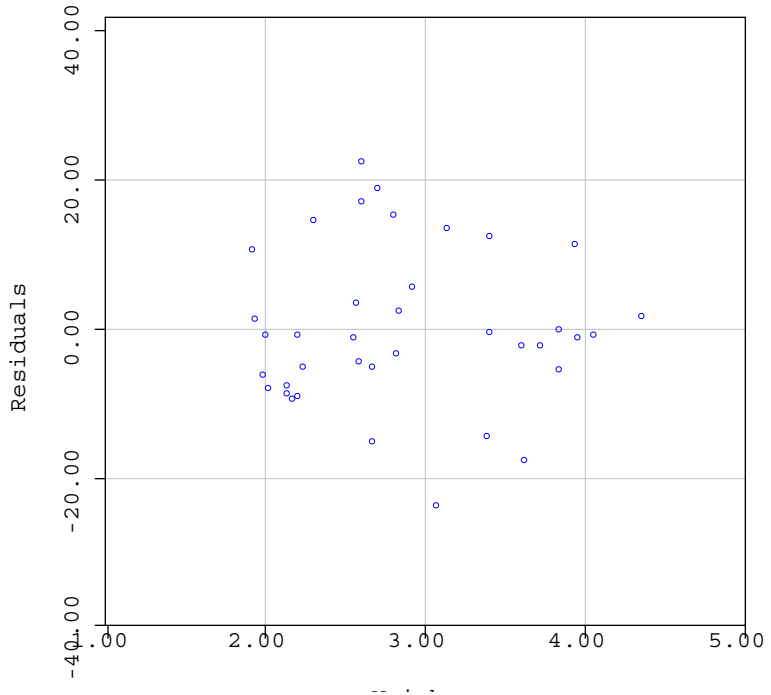


Aquí ponemos la variable predictora

- Ese gráfico debería mostrar una forma recta a lo largo del valor 0 en los residuales. Un ejemplo para una relación lineal sería el siguiente:

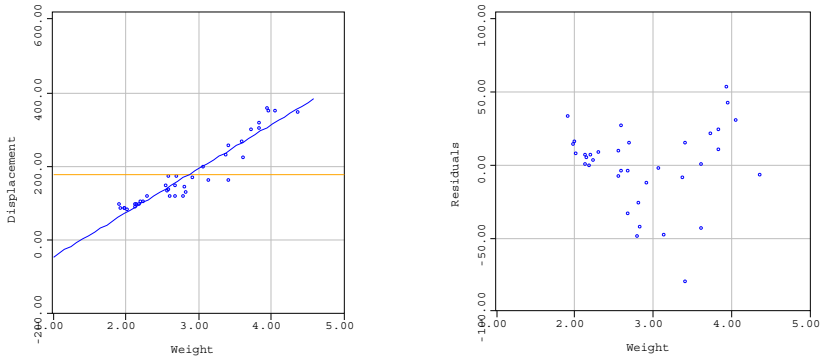


- En este caso, los residuales tienen la siguiente forma:

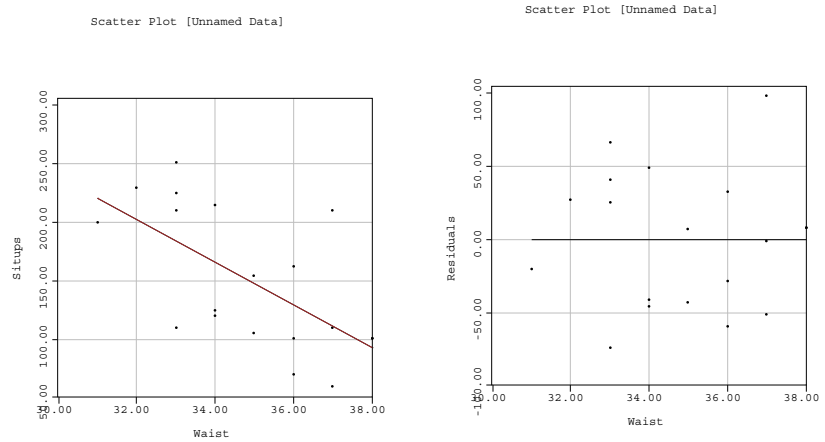


ACTIVIDADES

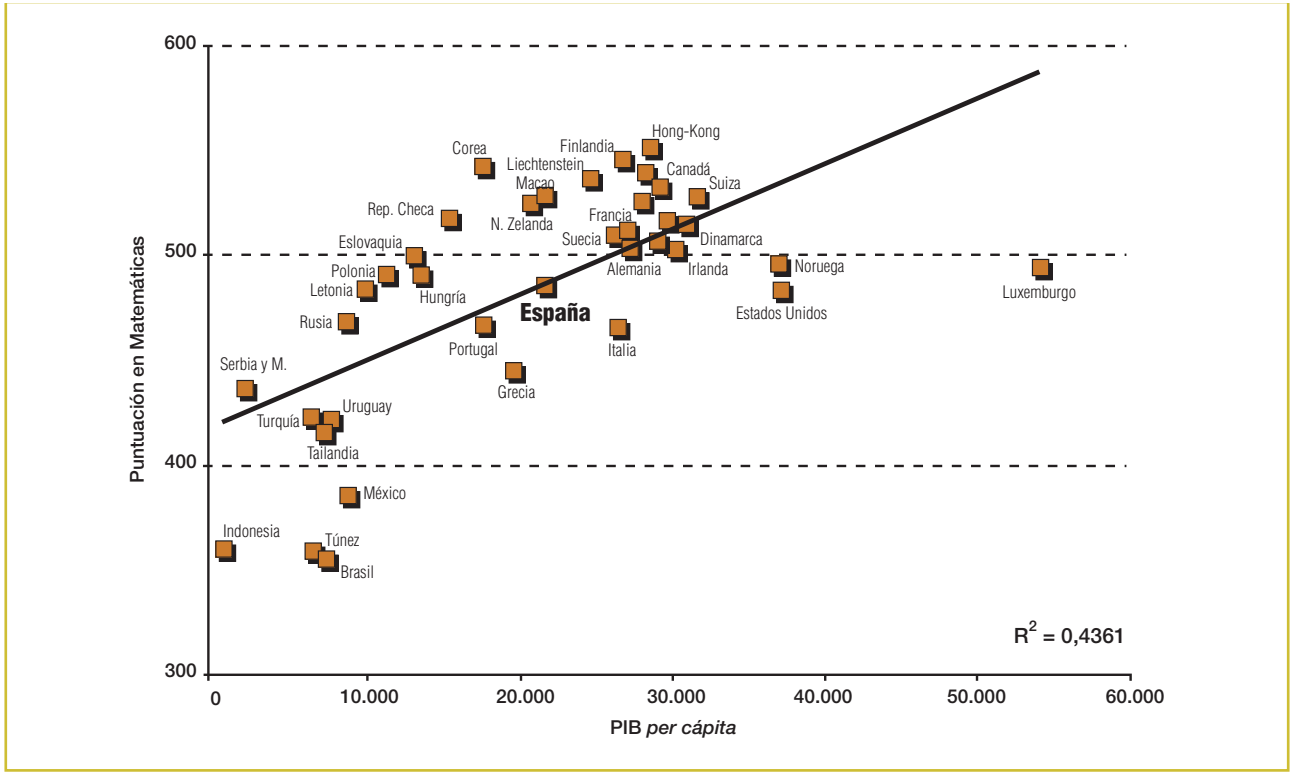
EJERCICIO 5.2.1 Indica si aprecias curvilinealidad en estos gráficos



EJERCICIO 5.2.2 Indica si aprecias curvilinealidad en la relación entre cintura (waist) y abdominales (situps)



EJERCICIO 5.2.3 ¿Dirías que hay curvilinearidad en este gráfico del informe PISA?



Fuente: *Banco Mundial y PISA*

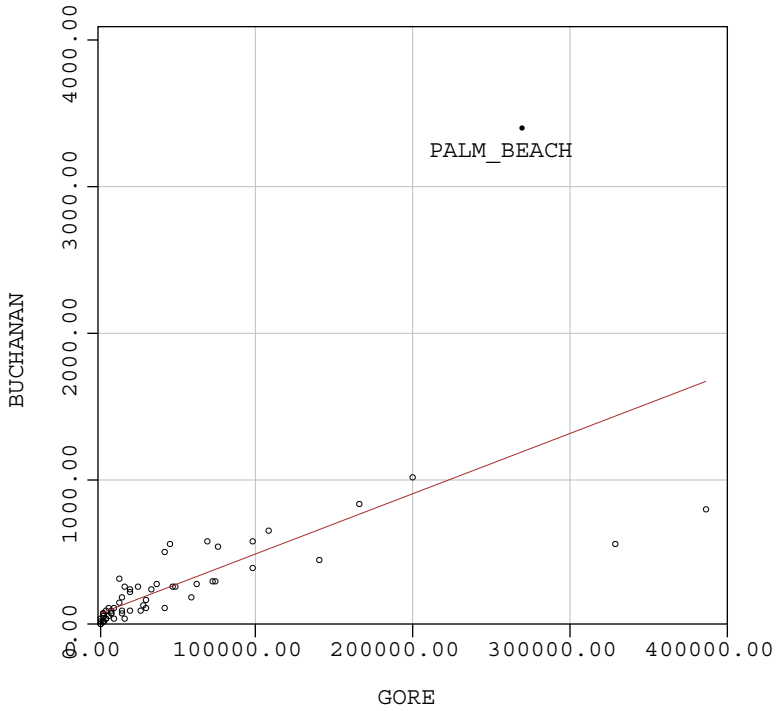
5.3.Evaluar residuales de gran tamaño

Evaluando valores extremos

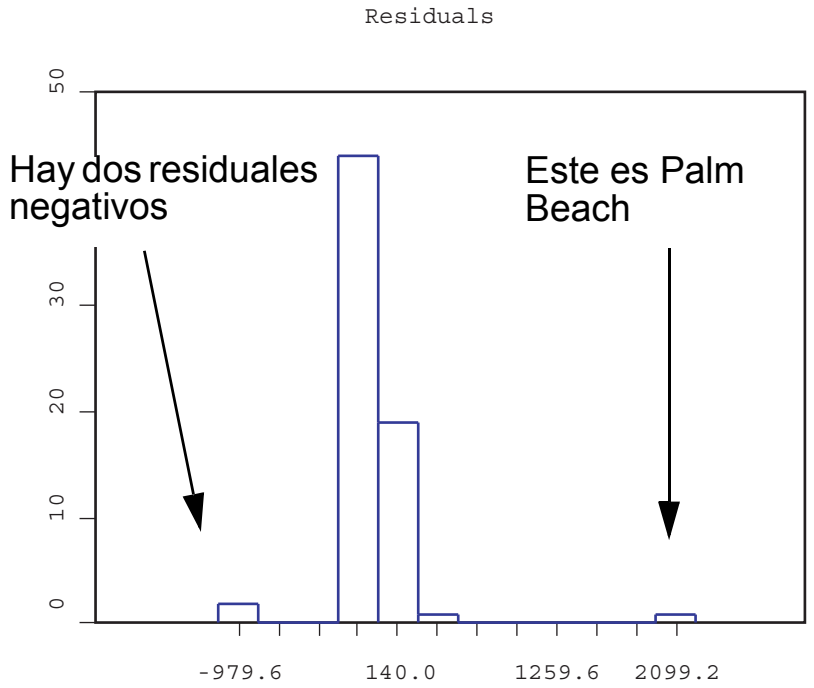
- En un análisis de regresión, algunos de los puntos pueden ajustar mucho peor que otros.
- Cuando los puntos que ajustan son unos pocos, y la diferencia es muy grande, esos puntos puede ser interesantes analizarlos con más detalle.
- Los valores extremos son valores que se dice que tienen más información que otros, por lo que resultan más interesantes que el resto.

- Ejemplo: En las elecciones del año 2000, hubo mucha controversia sobre las papeletas para votar en algunos condados en Florida. En esos condados, el diseño de la papeleta se supone que pudo llevar a que algunas personas que querían votar por Gore en realidad votaran a Buchanan. El condado más conflictivo fue PalmBeach

- Una forma de evaluar esto es ver el gráfico de votos de Gore v. Buchanan



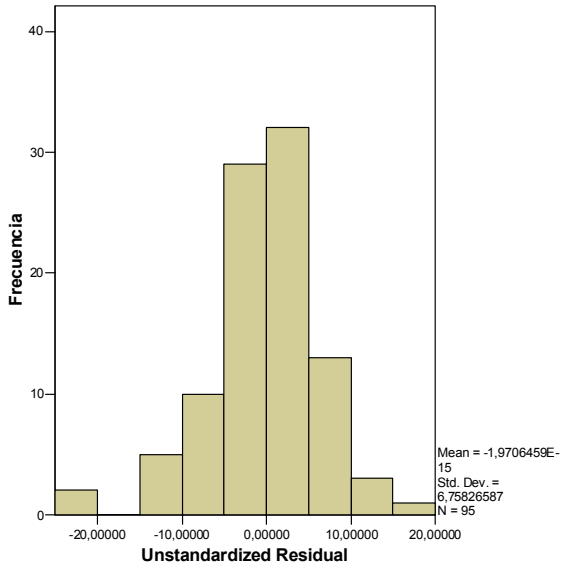
- Una forma de valorar los valores residuales extremos es hacer un histograma de éstos.



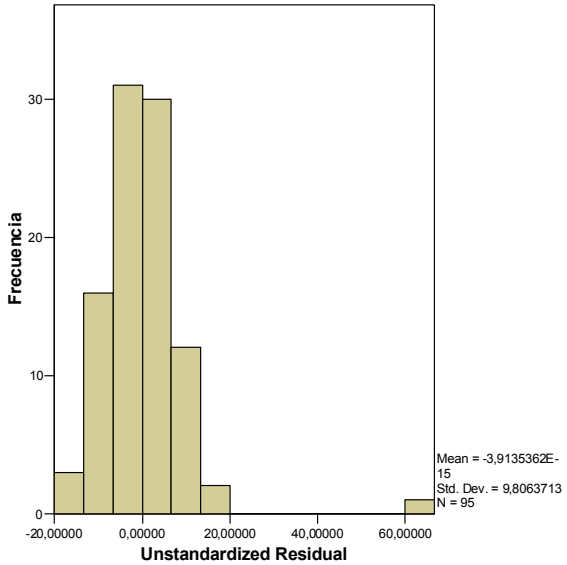
- ¿Qué hacemos con los residuales muy altos o bajos?
  - Los estudiamos por separado ya que a menudo los valores con residuales altos son ***más interesantes que los otros valores***
  - Damos el resultado para el resto de los datos después de haber excluido el valor residual ***pero informando de lo que hemos hecho.***

ACTIVIDADES

EJERCICIO 5.3.1 El siguiente gráfico muestra los residuales del análisis de regresión de la variable tanto por ciento de peso individual como predictora de la presión alta (presión sistólica) en los datos sobre lípidos. ¿Dirías que hay valores extremos?



EJERCICIO 5.3.2 El siguiente gráfico muestra los residuales del análisis de regresión de la variable tanto por ciento de peso individual como predictora de la presión baja (presión diastólica) en los datos sobre lípidos. ¿Dirías que hay valores extremos?

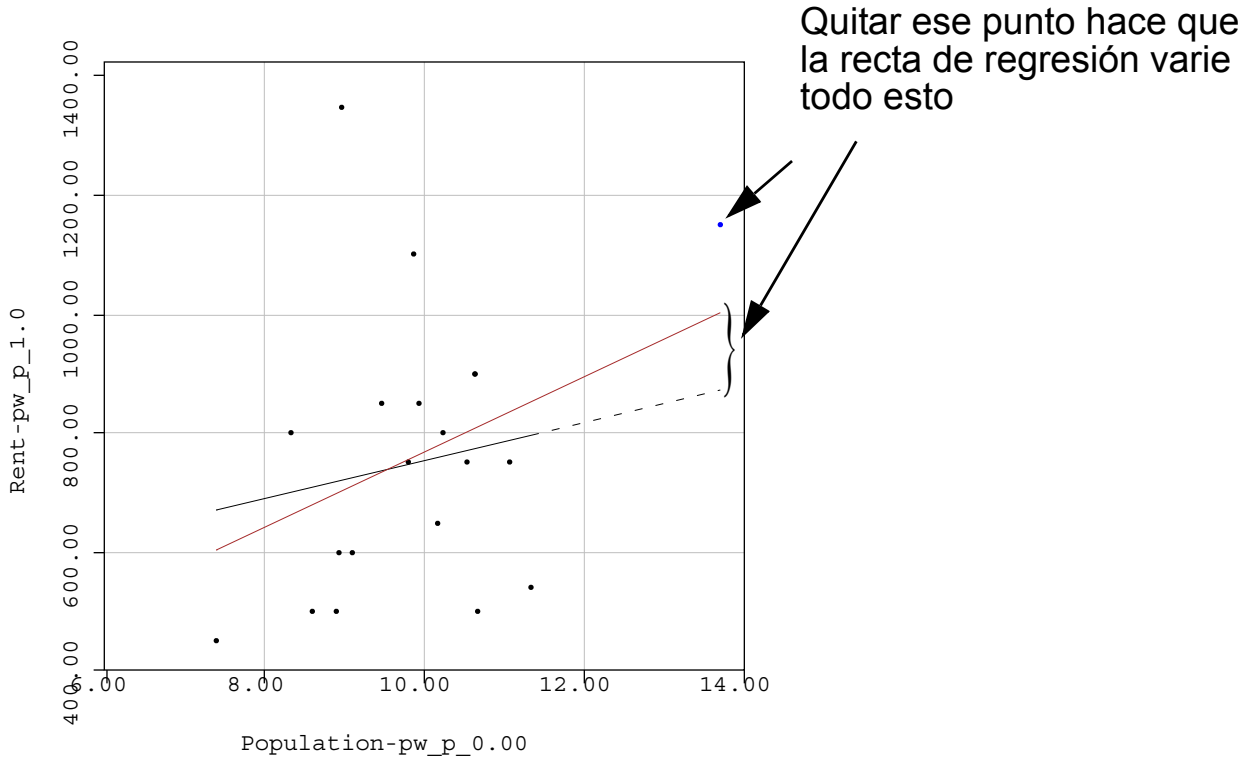


5.4.Evaluar puntos influyentes

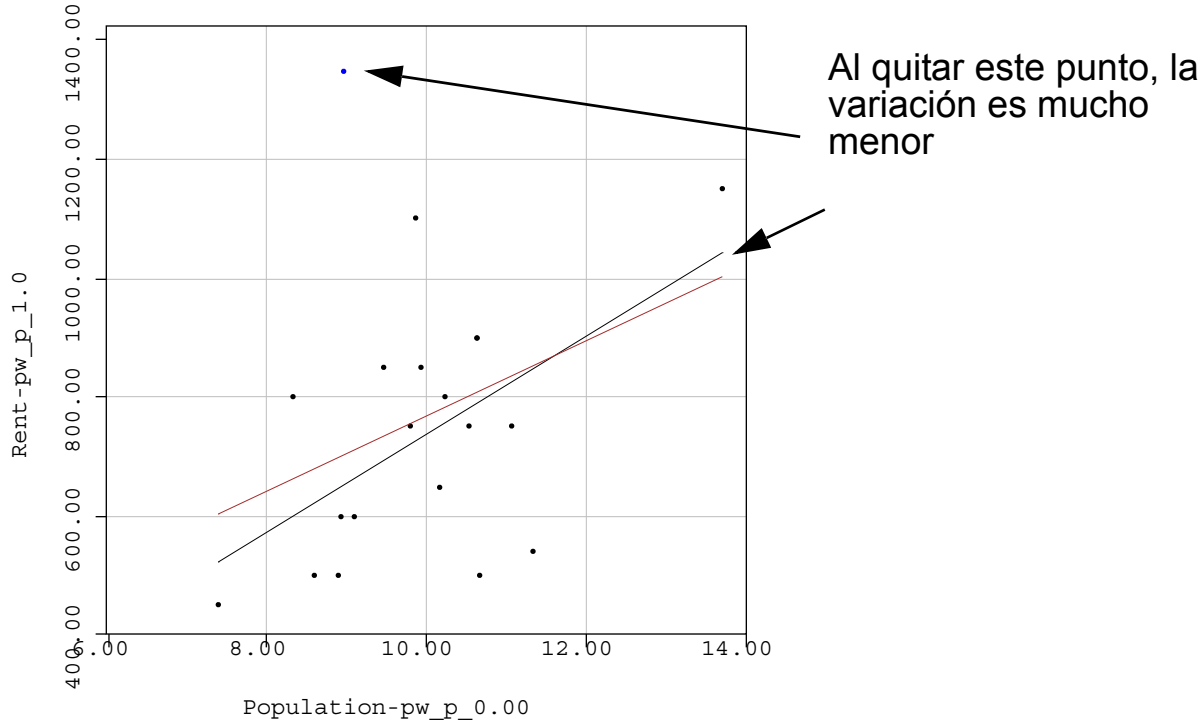
Todos los puntos deberían influir lo mismo

- Observaciones que tienen valores especialmente altos en ***la variable predictora*** pueden tener excesiva influencia sobre la regresión.
- Ejemplo: En un grupo de ciudades muy populares en Estados Unidos para jubilados tenemos la población (utilizaremos los logaritmos de la población por razones que no comentaremos) y el coste del alquiler de una casa.

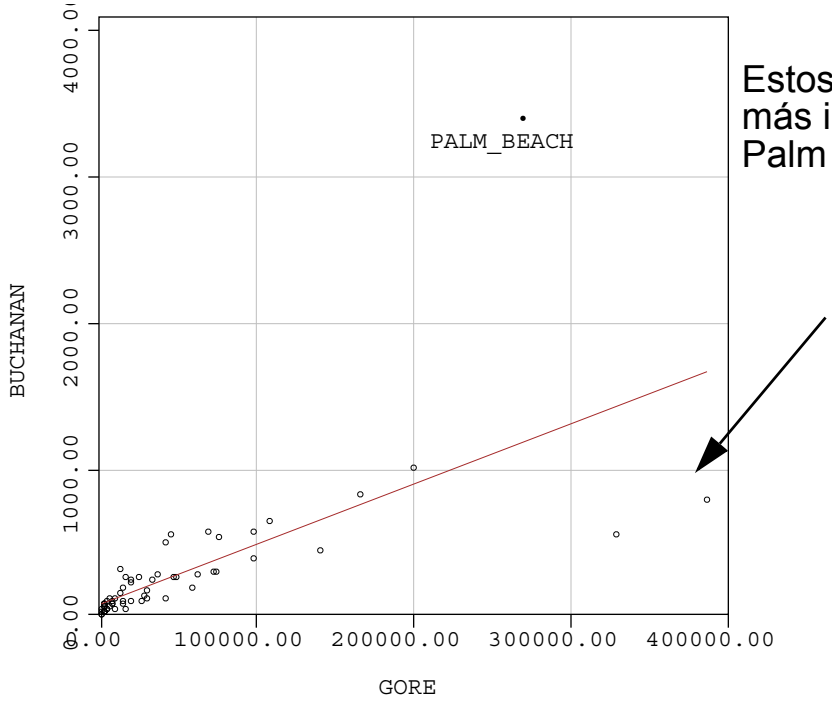
- En el diagrama de dispersión de estas dos variables hemos puesto dos líneas. Una ajusta a todos los datos, y la otra a todos menos Las Vegas, que es la ciudad con más habitantes del grupo de ciudades.



- Porque un punto sea extremo no tiene porque tener mucha influencia. Por ejemplo, si el punto que quitamos es el que está arriba:

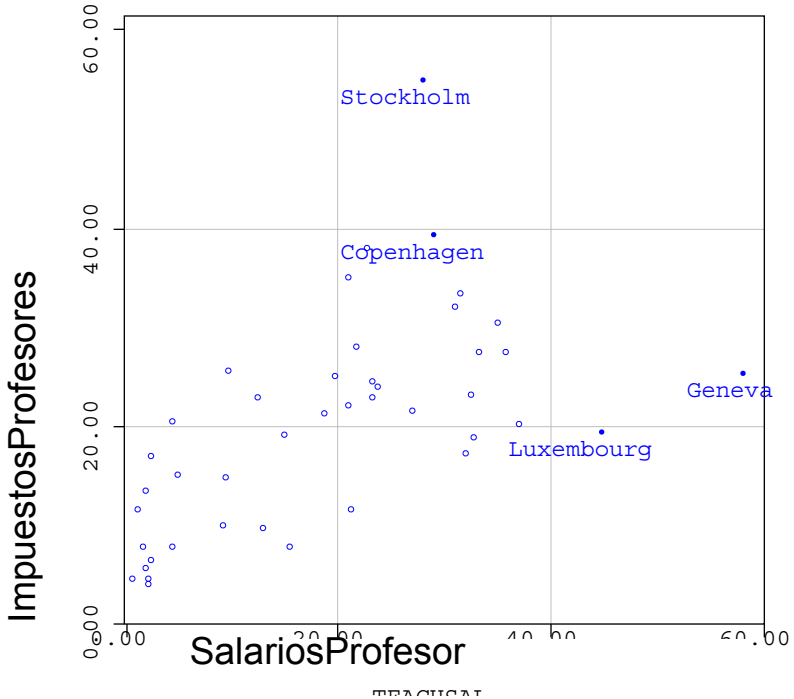


- Los puntos que tienen más influencia son los que destacan en la variable predictora, por la derecha o por la izquierda y no los que tienen un residual alto.
- En el ejemplo de las votaciones por Bush tendríamos lo siguiente:

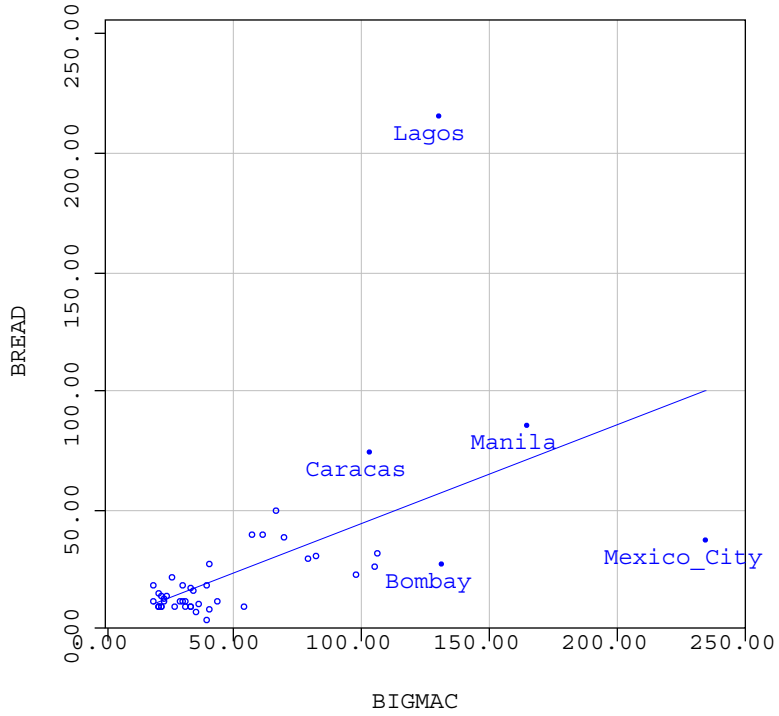


ACTIVIDADES

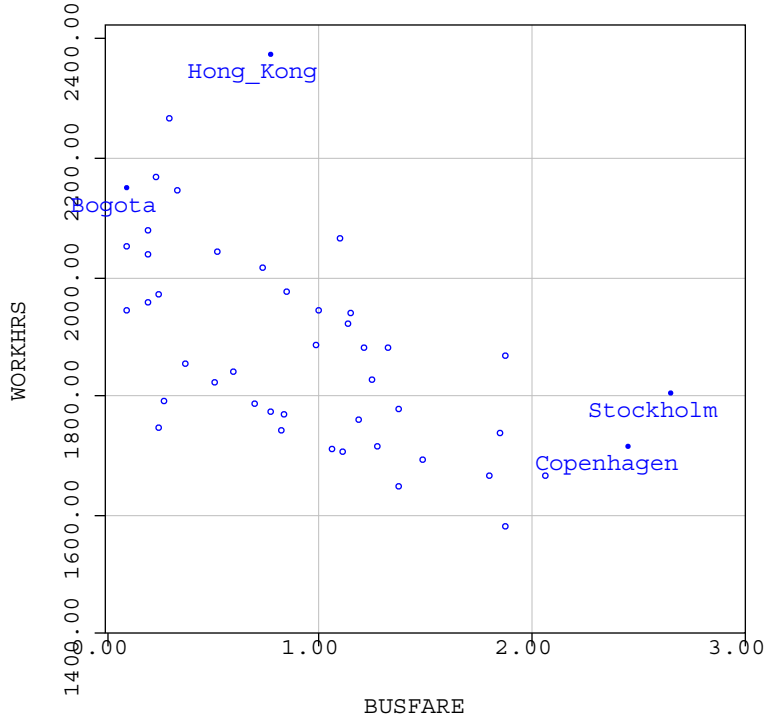
EJERCICIO 5.4.1 En el siguiente diagrama de dispersión se puede ver la relación entre la variable SalariosProfesores y la variable ImputestosProfesores. De las ciudades señaladas en el gráfico, ¿cuáles dirías que tienen más influencia?



EJERCICIO 5.4.2 Y de la regresión siguiente que utiliza BIGMAC como predictora y BREAD (pan) como predicha?



EJERCICIO 5.4.3 ¿Y en este caso? (BUSFARE=PRECIO DE UN RECORRIDO EN AUTOBÚS; WORKHRS=HORAS DE TRABAJO AL AÑO).

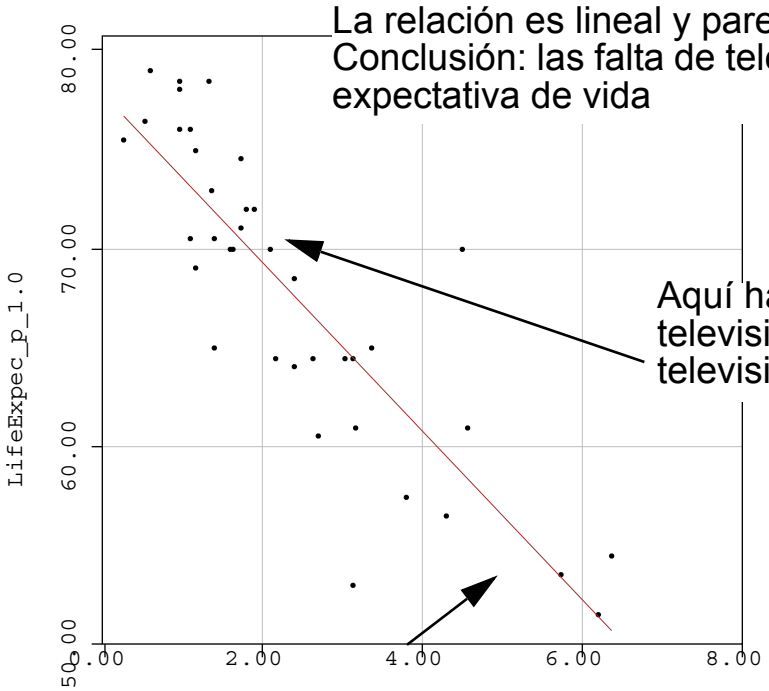


5.5.Pensar en variables subyacentes

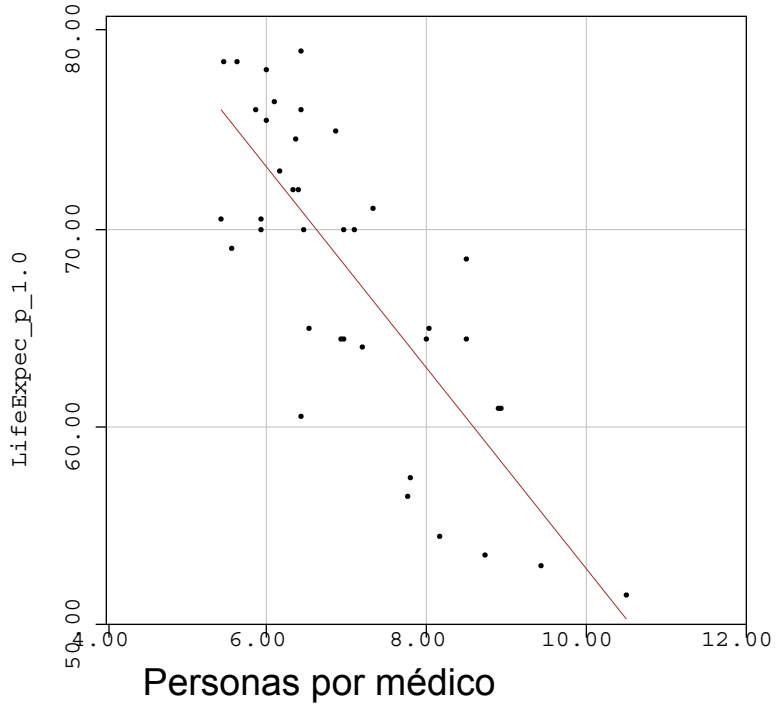
A veces las relaciones pueden ser muy sospechosas

- A veces, las relaciones entre dos variables pueden ser debidas a factores subyacentes o variables que denominamos intermedias.

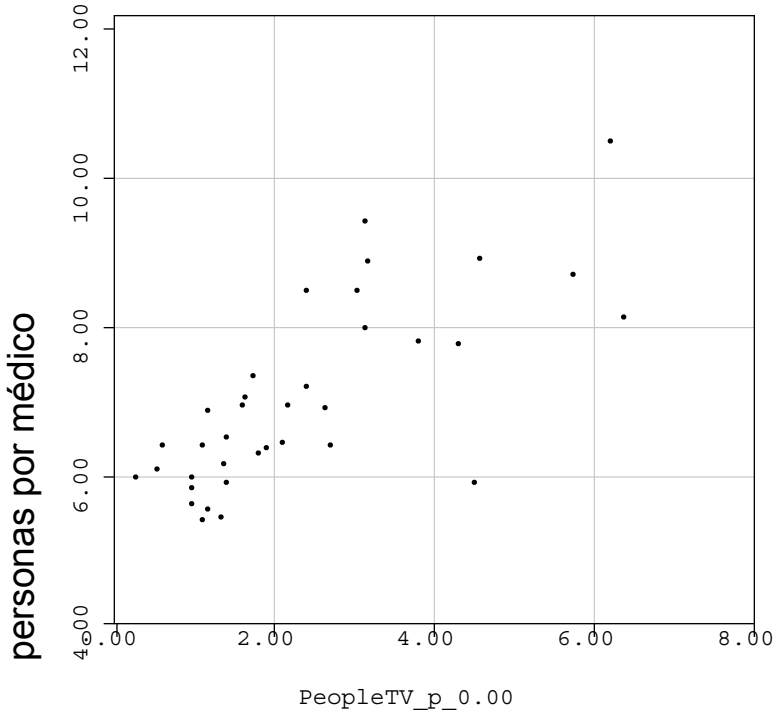
- Veamos el siguiente ejemplo. Tenemos la relación entre el número de personas por televisión (usaremos logaritmos) que hay en países del mundo y la expectativa de vida en ese país:



- No obstante, si pensamos un poco podemos ver que el número de personas por televisión es un indicador de la riqueza en un país, y que cuanta más riqueza, mejor sistema sanitario y más doctores tendremos en este sitio. Así, si hacemos esta regresión vemos que:



- En realidad, lo que pasa es que el número de médicos y el de televisiones está muy relacionado:

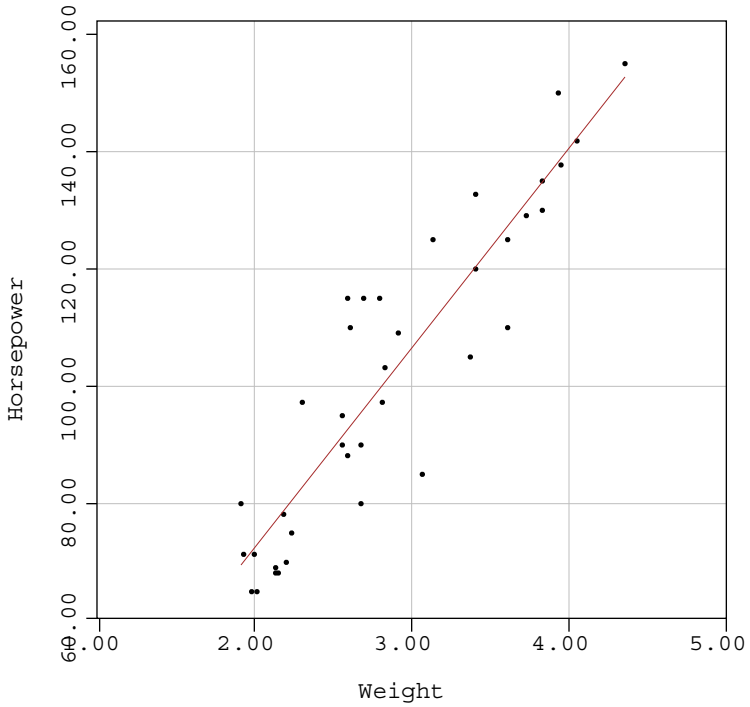


5.6.Soluciones al problema de curvilinealidad y los valores extremos

- Los problemas de curvilinearidad y de valores extremos que hemos visto antes pueden ser tratados con técnicas especiales. Aquí veremos unos métodos que permiten calcular coeficientes de correlación aunque la relación no sea lineal (o los datos vengan de rangos).

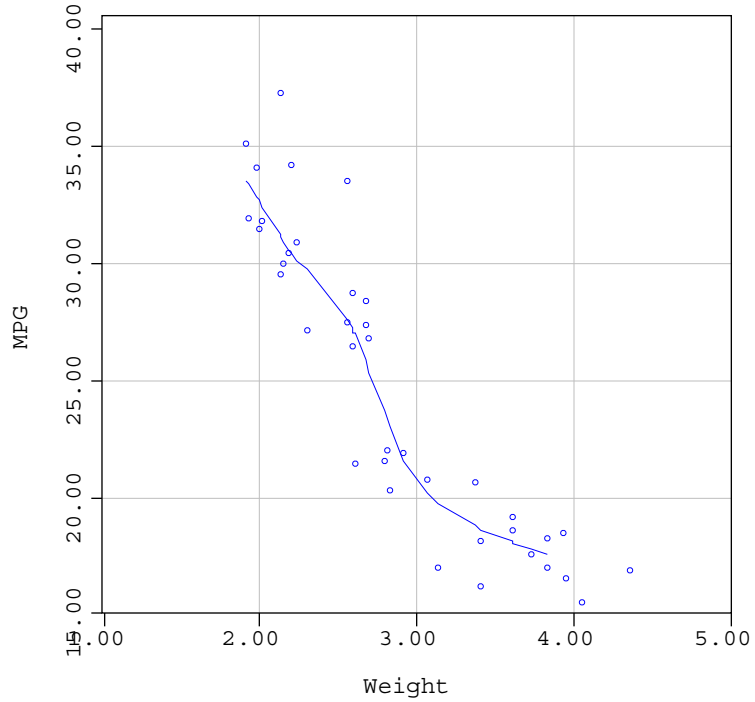
5.7.Soluciones : la tau de Kendall y la rho de Spearman

- La correlación de Pearson que vimos en la Sección 4.12. está diseñada para analizar problemas en los que las relaciones son líneas rectas. Por ejemplo:



- No obstante, cuando las relaciones son curvilíneas o hay valores extremos, la correlación de Pearson no es una buena indicación de la relación
- Un tipo especial de curvilinealidad es aquel en que las relaciones son siempre del mismo signo pero va variando de intensidad.
  - Estas relaciones se denominan monotónicas (y pueden ser crecientes o decrecientes)

- Un ejemplo de relación monotónica (decreciente) es la siguiente:



  - A medida que los coches tienen más peso recorren menos distancia pero ese efecto es más pronunciado con los coches pequeños que con los grandes

- ¿Cómo podemos medir esa asociación? Dos métodos que nos proporcionan una correlación para variables relacionadas monotónicamente son:
  - La tau de Kendall
  - La rho de Spearman

- Cálculo de la tau de Kendall
  - Veremos un ejemplo basado en la altura y el peso de un grupo de personas (este ejemplo está tomado de la Wikipedia).
  - El primer paso para calcular esos coeficientes es convertir las variables en rangos. En la Sección 3.25. ya vimos como convertir una variable en rangos.
  - En nuestro caso, los datos que tenemos son: los de la tabla de abajo. Fijaros que los datos están ordenados por la primera variable de modo que tenemos desde el más alto al más bajo

Table 19: Datos de rangos para un grupo de personas

Persona	A	B	C	D	E	F	G	H
Rango por Altura	1	2	3	4	5	6	7	8
Rango por Peso	3	4	1	2	5	7	8	6

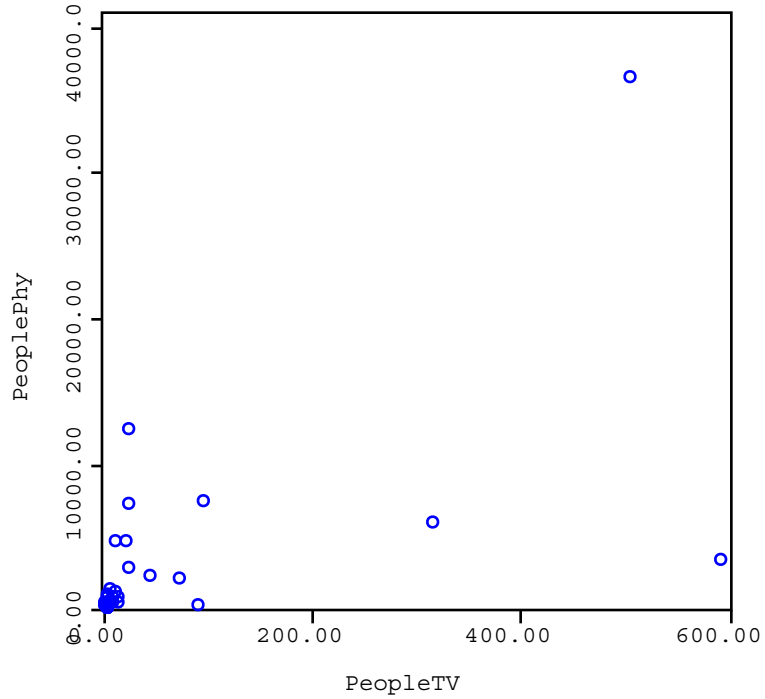
- Para hacer el cálculo vamos a la segunda variable (la que no está ordenada) y contamos para cada valor cuantos valores hay en esa misma variable (moviéndonos hacia la derecha) que son superiores a ese valor. Por ejemplo, el primer valor es 3 y hay 5 valores que están por encima de él (4, 5, 7, 8 y 6). El segundo valor es 4 y hay cuatro valores por encima (5, 7, 8 y 6)
- Haciéndolo para todos tenemos P = 5 + 4 + 5 + 4 + 3 + 1 + 0 + 0 = 22
- Ahora aplicamos la siguiente fórmula:

$$\tau = \frac{4P}{n(n-1)} - 1 = \frac{4(22)}{8(8-1)} - 1 = 0.57$$

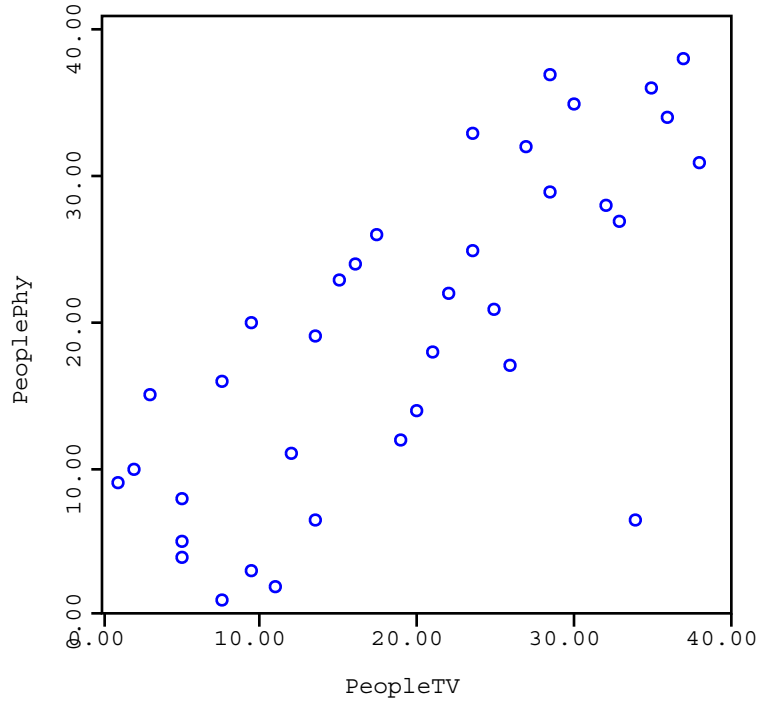
- Este coeficiente se interpreta como una correlación. Valores cercanos a 1 indican asociación entre los rangos, mientras que valores cercanos a -1 indican asociación inversa. Cero es no asociación.
- La fórmula que hemos usado aquí no tiene en cuenta que haya empates entre los rangos. Cuando eso ocurre hay otras fórmulas más especializadas que no veremos aquí.

- Cálculo de la rho de Spearman
  - Empezamos obteniendo los rangos
  - Luego aplicamos la fórmula de la correlación de Pearson.
  - La interpretación es como las otras correlaciones.

– Veamos un ejemplo. Este es el diagrama de dispersión para el número de personas por televisión y el número de personas por médico en una serie de países. En este caso calcular la correlación no parece una buena idea porque hay valores extremos.



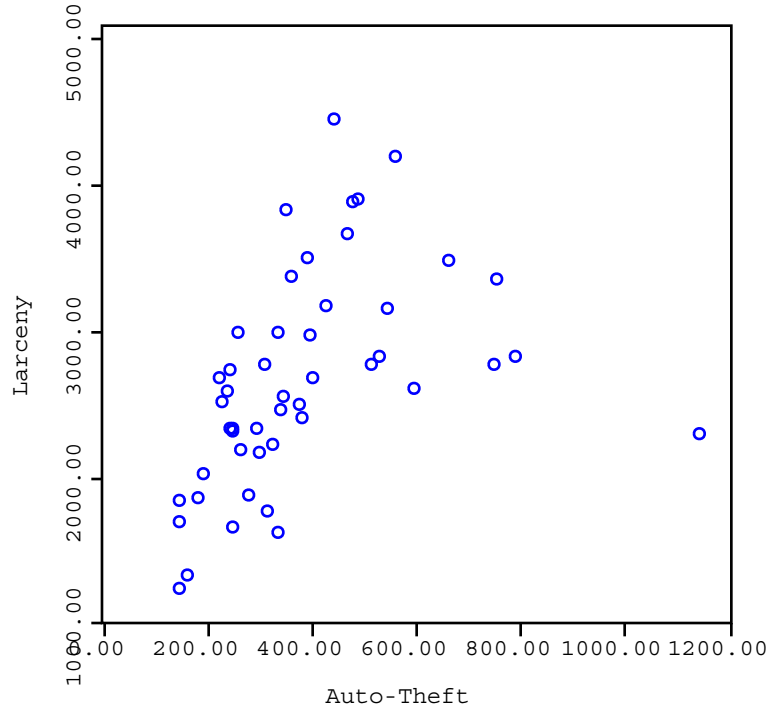
– Si calculamos los rangos de los países y hacemos el diagrama de dispersión tenemos lo siguiente:



- Ventajas e inconvenientes adicionales de la rho de Spearman y la tau de Kendall
  - Una ventaja importante es que son especialmente apropiados para datos que están en rangos. Por ejemplo, las escalas tipo Likert (en el que se pide a la gente que valore de 1 a 5 por ejemplo) pueden ser puestas en relación con estos coeficientes.
  - Estas correlaciones no se ven muy afectadas por valores que destacan mucho ya sean residuales o con influencia.
  - Estos dos coeficientes son métodos muy especializados. Si lo único que se quiere es calcular relaciones están bien, pero si se quiere hacer cosas más avanzadas ya no es posible.

ACTIVIDADES

EJERCICIO 5.7.1 Tenemos los datos de los crímenes en lugares de Estados Unidos. Viendo las variables Auto-Theft (robo de coches) y Larceny (Hurto) qué problema verías en calcular la correlación de Pearson?



5.8.Otros conceptos relacionados con correlaciones

- Covarianza
  - La fórmula de la covarianza es la siguiente

$$Cov(x,y) = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

  - La covarianza es una medida relacionada con la correlación

$$r_{(x,y)} = \frac{Cov(x,y)}{S_x S_y}$$

- La covarianza indica también relación entre variables pero a diferencia de la correlación no está limitada entre -1 y 1 sino que puede tomar todos los valores
- En la práctica, la correlación es más útil aunque si todas las variables comparadas están en la misma escala (por ejemplo cuando son respuestas a preguntas de cuestionario) puede ser interesante
- También forma parte de fórmulas más complejas

- El coeficiente de regresión para puntuaciones típicas
  - En su momento vimos que la fórmula para una recta de regresión es la siguiente:

$$\hat{y} = a + bx$$

  - Si cambiamos x e y a puntuaciones típicas y calculamos la regresión tenemos que:

$$\hat{z}_{(y)} = rz_{(x)}$$

  - Es decir, la pendiente en puntuaciones típicas es el coeficiente de correlación (y la intercepta 0)

- Matriz de correlaciones/covarianzas
  - Ya vimos en clase de prácticas que podemos disponer varias correlaciones en una tabla de tamaño cuadrado. Por ejemplo

	Aficiones y Hobbies	Música clásica	Visitó museos o galerías en el último año
Aficiones y Hobbies	1	,232	,165
Música clásica	,232	1	,371
Visitó museos o galerías en el último año	,165	,371	1

  - Esta tabla tiene un significado especial en matemáticas: se llama matriz y en concreto las de correlaciones son simétricas (la parte de abajo es igual a la de arriba) con diagonal igual a 1
  - Las operaciones sobre matrices son muy importantes para el cálculo matemático en estadística (aunque son bastante pesadas a mano)

Parte VI
Más de dos variables numéricas

6.1.Introducción

- Los análisis que hemos visto anteriormente son aplicables a situaciones en las que tenemos más de dos variables. Hay dos casos principales:
  - Cuando tenemos más de una variable predictora con lo que podemos mejorar la predicción. Esto se llama análisis de regresión múltiple
  - Cuando queremos entender las interrelaciones entre las variables de una matriz de correlaciones más allá de considerar sólo dos variables cada vez. Aquí veremos el análisis de componentes principales

6.2.Análisis de regresión múltiple

- Usaremos el ejemplo de predicción del sueldo a partir de otras variables. El archivo de datos está en la página del curso y tiene dos candidatas: educación necesaria y prestigio de las profesiones.

- Análisis para educación

Correlación

Correlación al cuadrado

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,725[a]	,525	,514	17,03688

a. Variables predictoras: (Constante), Educación

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	10,603	5,198		2,040	,048
	Educación	,595	,086	,725	6,893	,000

a. Variable dependiente: Salario

Coeficientes de la regresión

- Análisis para prestigio

Correlación

Correlación al cuadrado

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,838[a]	,702	,695	13,49518

a. Variables predictoras: (Constante), Prestigio

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	10,884	3,678		2,959	,005
	Prestigio	,65	,065	,838	10,062	,000

a. Variable dependiente: Salario

Coeficientes de la regresión

- Mirados en conjunto, los análisis anteriores muestran que tanto el prestigio como la educación son buenos predictores del salario
  - Notar que todavía no hemos entrado en la fase de diagnóstico y que en estos ejemplos hay varios valores extremos
  - El factor educación explica el 52.5% de la varianza mientras que el prestigio explica el 70%
- Ahora bien, si combinaramos las dos variables, qué cantidad de varianza podemos aspirar a explicar
  - Desde luego, 52.5+70=132.5% es imposible!! No se puede explicar más del 100% de la varianza

- El resultado de la regresión múltiple se muestra a continuación

Correlación

Correlación al cuadrado

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,838[a]	,702	,688	13,64519

a. Variables predictoras: (Constante), Prestigio, Educación

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	10,426	4,164		2,504	,016
	Educación	,032	,132	,039	,244	,808
	Prestigio	,624	,125	,804	5,003	,000

a. Variable dependiente: Salario

Coeficientes de la regresión

  - En este caso la fórmula calculada incluye Educación y Prestigio simultaneamente para predecir el salario

- No obstante, vemos que la proporción de varianza explicada es practicamente la misma que la obtenida mediante el Prestigio sólo. Esto significa que en este caso no hemos ganado nada al utilizar dos variables predictoras en lugar de sólo el prestigio a la hora de predecir el sueldo.
- Los mejores predictores en un análisis de regresión múltiple son los que están poco relacionados entre sí pero bastante relacionados con la variable predicha (en este caso la Educación y el Prestigio están muy relacionados entre sí-su correlación es 0.852)
- Una forma de valorar la aportación individual de cada variable es mirar los coeficientes tipificados (Beta). Estos valores son los coeficientes de regresión para

variables que han sido pasadas a típicas. Estos valores pueden ser comparados entre sí para ver qué variables tienen más influencia sobre la variable dependiente. Hacer esta comparación para los coeficientes no estandarizados no es correcto

ACTIVIDADES

EJERCICIO 6.2.1 A un grupo de niños de edades comprendidas entre 4 y 7.5 se les hizo una serie de pruebas para evaluar su capacidad de lectura. Esta capacidad fue puesta en función de la edad y y la inteligencia general (medida mediante un cuestionario). Un análisis de regresión múltiple con ambos factores produjo los siguientes resultados.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,906[a]	,821	,800	,30739

a. Variables predictoras: (Constante), IQ, age

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-,703	1,247		-,564	,580
	age	,584	,067	,912	8,711	,000
	IQ	,036	,011	,331	3,165	,006

a. Variable dependiente: reading ability

6.3.Análisis de componentes principales

- El análisis de componentes principales intenta resumir la información de matrices de correlación en una serie de componentes o factores
- Esta técnica permite realizar gráficos para visualizar esa información más fácilmente

Ejemplo

En el ejemplo sobre preferencias en gustos musicales y otras caracterís-ticas encontramos que ciertos gustos musicales en los encuestados estaban correlacionados con otros. Una forma de agrupar esos gustos musicales es utilizar el análisis de componentes principales. El resultado principal

de este análisis es el siguiente:

Correlaciones de los factores con las variables

Matriz de componentes[a]

	Componente										
	1	2	3	4	5	6	7	8	9	10	11
Música de bigband	,713	-,124	-,079	-,141	,294	,466	-,016	-,138	-,071	-,334	,104
Música country western	,144	-,557	,600	,022	,315	-,075	,244	,367	,039	,024	,086
Música Blues o Rhythm & Blu	,531	,362	,333	-,497	-,123	-,102	,021	,029	-,435	,107	,040
Musicales de Broadway	,743	-,033	-,266	,073	,252	,214	-,232	,131	,029	,406	-,158
Música clásica	,741	,073	-,334	,243	-,168	-,141	,140	,008	,074	,103	,438
Música popular	,625	-,341	,091	,257	-,383	-,121	-,376	,231	-,059	-,226	-,096
Música Jazz	,526	,491	,131	-,485	-,079	-,066	,020	,129	,436	-,089	-,072
Ópera	,712	,061	-,228	,267	,089	-,264	,414	-,105	-,074	-,098	-,300
Música Rap	,087	,592	,388	,376	,417	-,269	-,286	-,111	,003	-,071	,072
Música Heavy Metal	-,018	,549	,404	,462	-,262	,445	,194	,123	-,009	,035	-,041
Música bluegrass	,426	-,430	,584	,017	-,179	,017	-,019	-,465	,140	,152	-,024

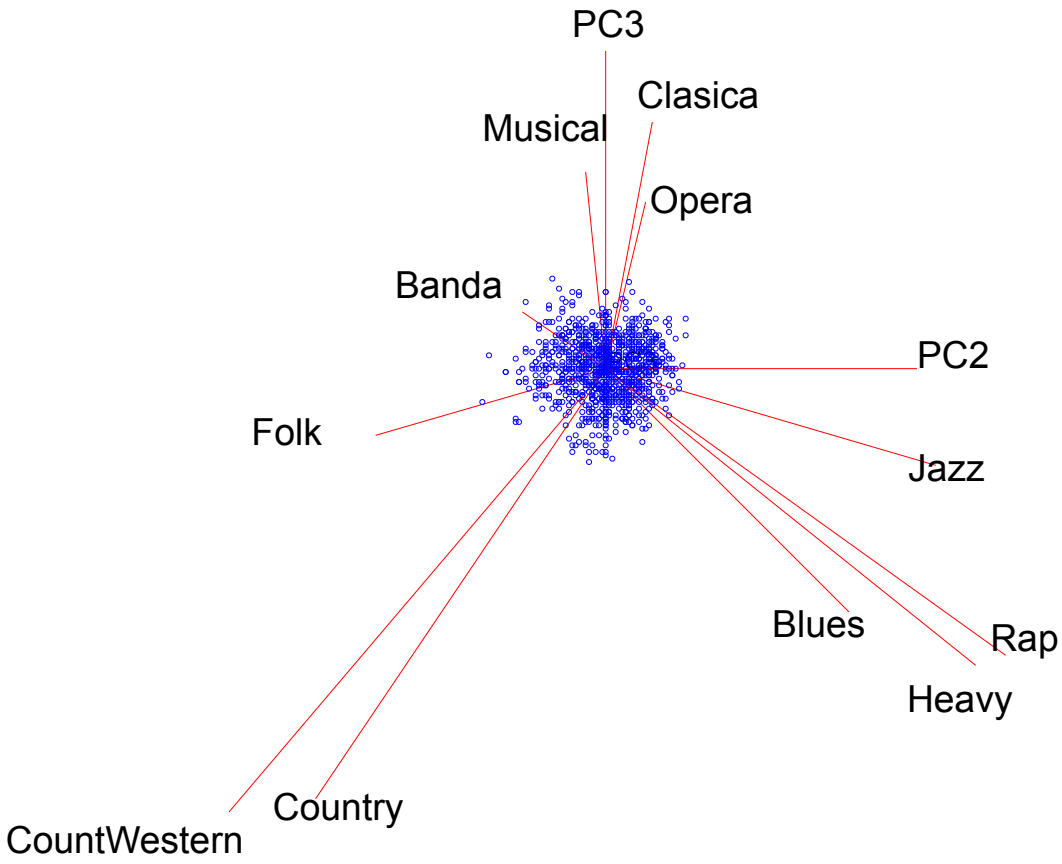
Método de extracción: Análisis de componentes principales.

a. 11 componentes extraídos

- Por ejemplo, en el componente 2 puntuan amantes del Jazz, del Rap y del Heavy que no les gusta el country western ni el bluegrass

- En un análisis de componentes principales, los primeros factores explican más varianza que los siguientes
  - A menudo, se seleccionan unos pocos de los primeros y se ignoran el resto
  - No obstante, a veces el más interesante no es el primero. Puede ocurrir que los resultados más interesantes empiezan en el segundo componente
  - En nuestro caso, los componentes 2 y 3 ofrecen, en mi opinión, los resultados más interesantes

- Representación gráfica del resultado



- En esta representación gráfica, las correlaciones altas y positivas se convierten en ángulos agudos, las altas y negativas en ángulos llanos, y las nulas en ángulos rectos.
- En el ejemplo vemos que PC2 está relacionado positivamente con el Jazz, el Blues, El Rap y el Heavy y negativamente con el Country y el Countrywestern
- El PC3 está relacionado positivamente con la música Clásica, la Opera y el Musical mientras que está relacionado negativamente con la música Country y CountryWestern
- Algunos tipos de música tienen posiciones intermedias (Jazz, Folk, Banda)