Estadística para Periodistas

Andrés M. Alonso

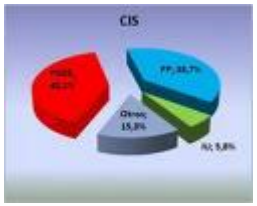
Departamento de Estadística
Universidad Carlos III de Madrid

E. Mail: andres.alonso@uc3m.es

Web: www.est.uc3m.es/amalonso

http://www.est.uc3m.es/amalonso/esp/EAP2014.htm

Motivación: El uso de la estadística



Encuestas

Estimación del desempleo

Predicciones económicas

Toma de decisiones

Bibliografía

- Gonick, L. y Smith, W. (2010) *La Estadística en Comic*. Editorial Zendrera Zariquiey, Barcelona.
- Jauset. J.A. (2000) La investigación de audiencias en televisión Fundamentos estadísticos, Editorial Paidós, Barcelona.
- Jauset, J.A. (2007) *Estadística para periodistas, publicitarios y comunicadores*, Editorial UOC, Barcelona
- Peña, D. y Romo, J. (2009) *Introducción a la Estadística para las Ciencias Sociales*, Editorial McGraw-Hill, Madrid.
- Pérez, C. (2010) *Estadística Aplicada a través de Excel*, Editorial Prentice-Hall, Madrid.
- Portilla, I. (2004) *Estadística descriptiva para comunicadores* Editorial EUNSA, Pamplona.
- Wimmer, R. y Dominick, J. (2001) *Introducción a la investigación en medios masivos de comunicación*, International Thomson Editores.

Blografía

- Estadística por todas partes
- Un mar de datos
- Estadística para todos
- El periodista y los números
- Malaprensa sobre errores estadísticos en la prensa

Concepto y usos de la estadística

a) ¿Qué es la estadística?

b) Precauciones ante la estadística.

c) ¿Para qué sirve la estadística?

a: ¿Qué es la estadística?

La RAE define la estadística así:

estadística.
(Del al. *Statistik*).

1. f. Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.

2. f. Conjunto de estos datos.

3. f. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

b: Precauciones ante la estadística

En muchos casos es posible utilizar la estadística para influir en el público.

Respetabilidad o soporte estadístico del discurso.

Ninguno de los gráficos es incorrecto

b: Precauciones ante la estadística

Estos periodistas necesitan
un curso de estadística

Tres portadas

El Mundo, 21 de febrero de 2005, tras la aprobación en referendum de la Constitución Europea por un 76% de los votantes (32% del censo):
Rotunda victoria del 'sí' a la Constitución con una participación baja pero aceptable

El Mundo, 19 de junio de 2006, tras la aprobación en referendum del Estatuto de Cataluña por un 74% de los votantes (36% del censo):
La mayoría de los catalanes da la espalda al Estatuto que les define como 'nación'

El Mundo, 19 de febrero de 2007, tras la aprobación en referendum del Estatuto de Andalucía por un 87,5% de los votantes (31% del censo):
Sólo el 31% de los andaluces refrenda su 'realidad nacional'

Fuente: Malaprensa

Ejercicio

¿Qué piensas tu?

La Vanguardia (13/12/2009)

“El 'sí' gana en las consultas soberanistas con el 94,9%”

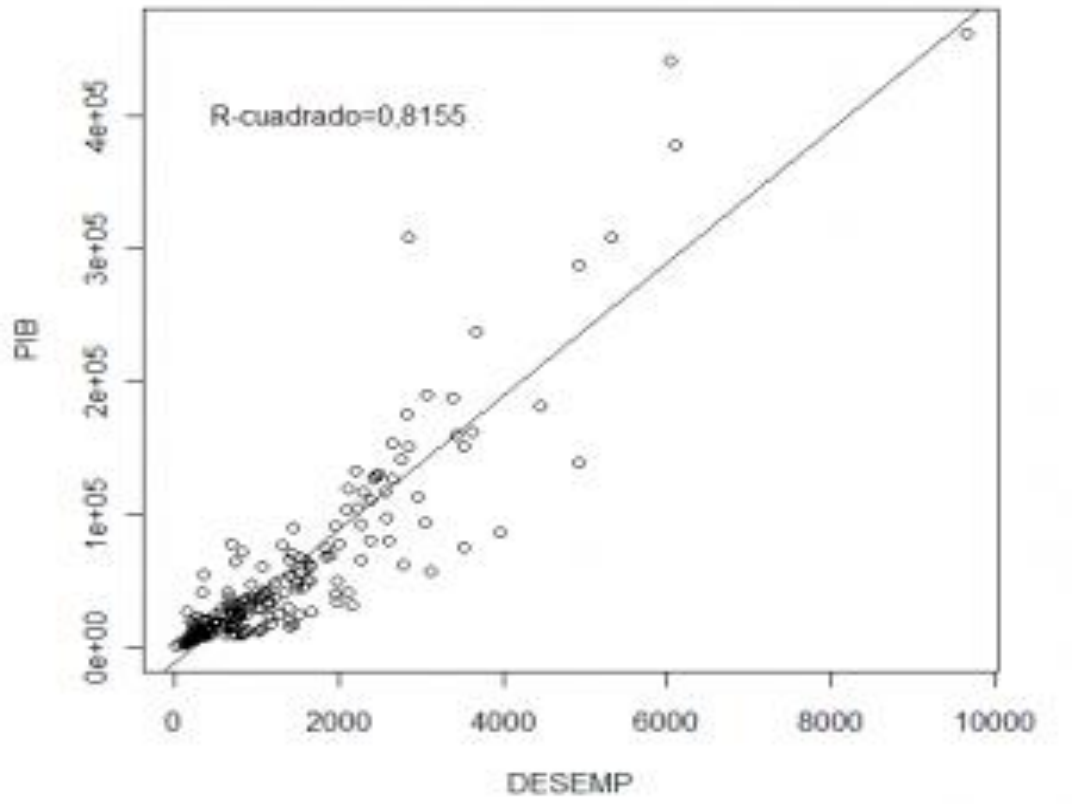
El recuento final destapó el 3,2% de votos en contra, un 1,6% de votos en blanco y un 0,3% nulos. La Coordinadora calificó la jornada de "heroicidad" por los medios disponibles …

ABC (13/12/2009)

“Las consultas independentistas pinchan con una participación por debajo del 30%”

Según los datos ofrecidos por los organizadores, en los 166 municipios convocados –con un censo de unas 700.000 personas mayores de 16 años, entre españoles y extranjeros- sólo votaron 200.000, rozando el 30%.

Datos para 187 regiones europeas



Resulta que cuánto mayor es el número de personas que no tienen trabajo es mayor el PIB.

c: ¿Para qué sirve la estadística?

- **Recogida y resumen de información**
- **Ilustración de relaciones entre distintas variables**
- **Cambios en una variable en el tiempo**

- **Estimación de las características de una población a través de una muestra.**
- **Previsión**

Estadística Descriptiva **Estadística Inferencial**

Cuatro cosas que se deben saber

i. Media versus mediana

ii. Pictogramas correctos

iii. Frecuencias absolutas, relativas y condicionadas

iv. Intervalos (*horquillas*) de confianza

i) Media versus mediana

Veamos que dice la RAE:

promedio.
(Del lat. *pro medio*).
1. m. Punto en que algo se divide por mitad o casi por la mitad.
2. m. **término medio** (‖ cantidad igual o más próxima a la media aritmética).

y Wikipedia:

En matemáticas y estadística, la **media aritmética** (también llamada **promedio** o simplemente **media**) de un conjunto finito de números es el valor característico de una serie de datos cuantitativos objeto de estudio … se obtiene a partir de la suma de todos sus valores dividida entre el número de sumandos.

En Excel se calcula como =promedio()

i) Media versus mediana

Supongamos que tenemos una empresa con 20 empleados y 5 directivos. Por simplificar, supongamos que todos los empleados tienen un salario de 1000 y que los directivos tiene un salario 10 veces mayor.

La **media** es (20x1000 + 5x10000)/25 = 2800

Ese valor no está en el medio ni divide por la mitad y es cuestionable como valor característico de los salarios de esa empresa

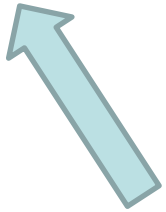
Podría ser aún peor, añadimos el salario asignado al propietario de la empresa …

Valor atípico u *outlier*

i) Media versus mediana

Si la media no es la mitad ni parece ser un valor característico (representativo) en esta situación ¿qué usar?

mediano, na.
(Del lat. *mediānus*, del medio).
9. f. *Mat.* Elemento de una serie ordenada de valores crecientes de forma que la divide en dos partes iguales, superiores e inferiores a él.

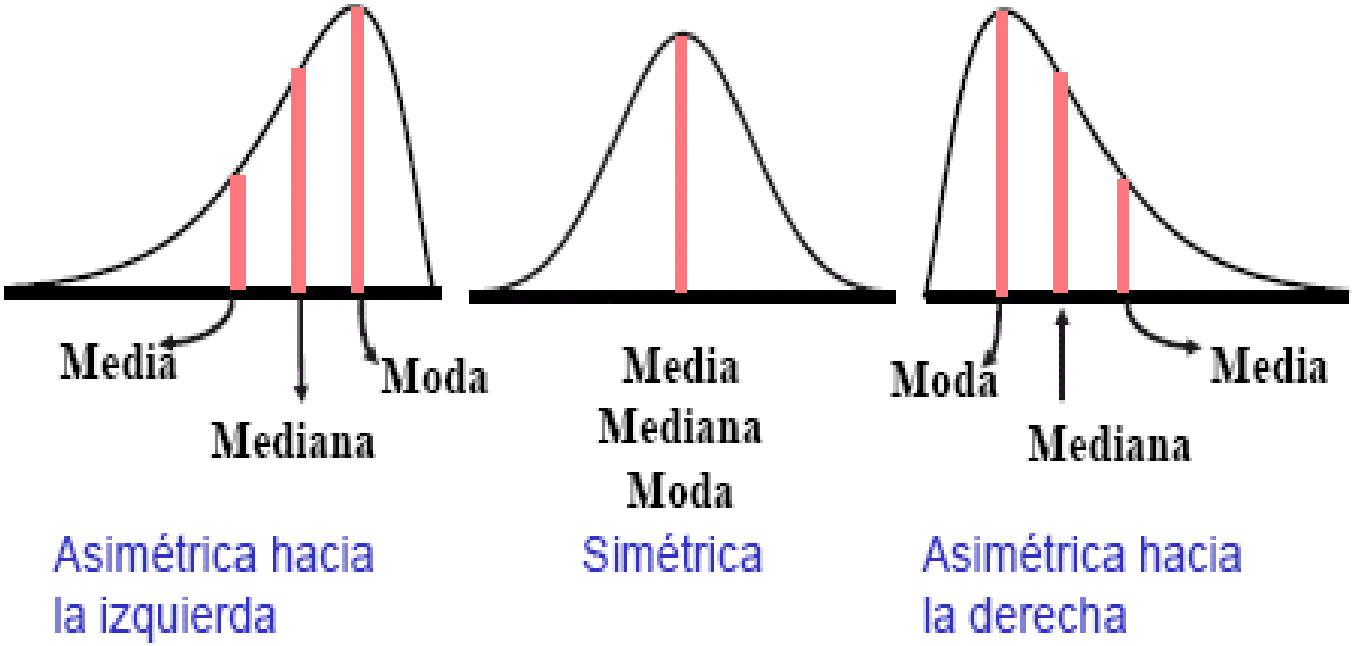


La propia definición garantiza que el valor divide a la muestra (o la población) en dos partes iguales.

En Excel se calcula como =mediana()

i) Media versus mediana

¿Cuándo usar media y cuándo mediana?



Adicionalmente, la media puede considerarse “precisa” si el coeficiente de variación, $CV = s/|\bar{x}|$, es pequeño.

Diagrama de caja

Espero haberos convencido de la utilidad de la mediana para describir la posición central de una variable.

¿Cómo complementarla?

Con los cuartiles pero tantos números

(se) pueden perder a los lectores.

Mejor un gráfico ¿cuál?

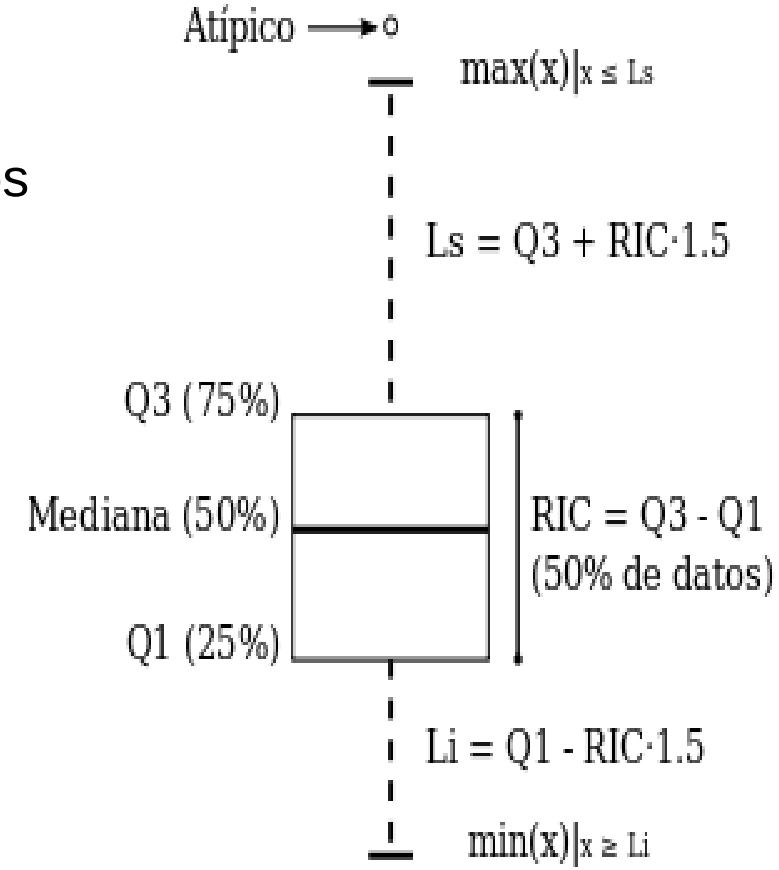


Diagrama de caja

CIS

Estudio nº3041. BARÓMETRO DE OCTUBRE 2014 **Octubre 2014**

Pregunta 16

Como Ud. sabe, en España hay distintos partidos o coaliciones políticas a las que puede votar en unas elecciones. Me gustaría que me dijera cuál es la probabilidad de que Ud. vote a cada uno de los que le voy a mencionar, utilizando para ello una escala de 0 a 10, sabiendo que el 0 significa que 'con toda seguridad, no le votaría nunca' y el 10 significa que 'con toda seguridad, le votaría siempre'.

Pregunta 26

Y, utilizando esa misma escala, por favor dígame dónde colocaría Ud. a cada uno de los siguientes partidos o formaciones políticas.

Escala de la pregunta 26: Izquierda 1 - Derecha 10.

Veamos primero las medias calculadas para las respuestas a las preguntas 16 y 26.

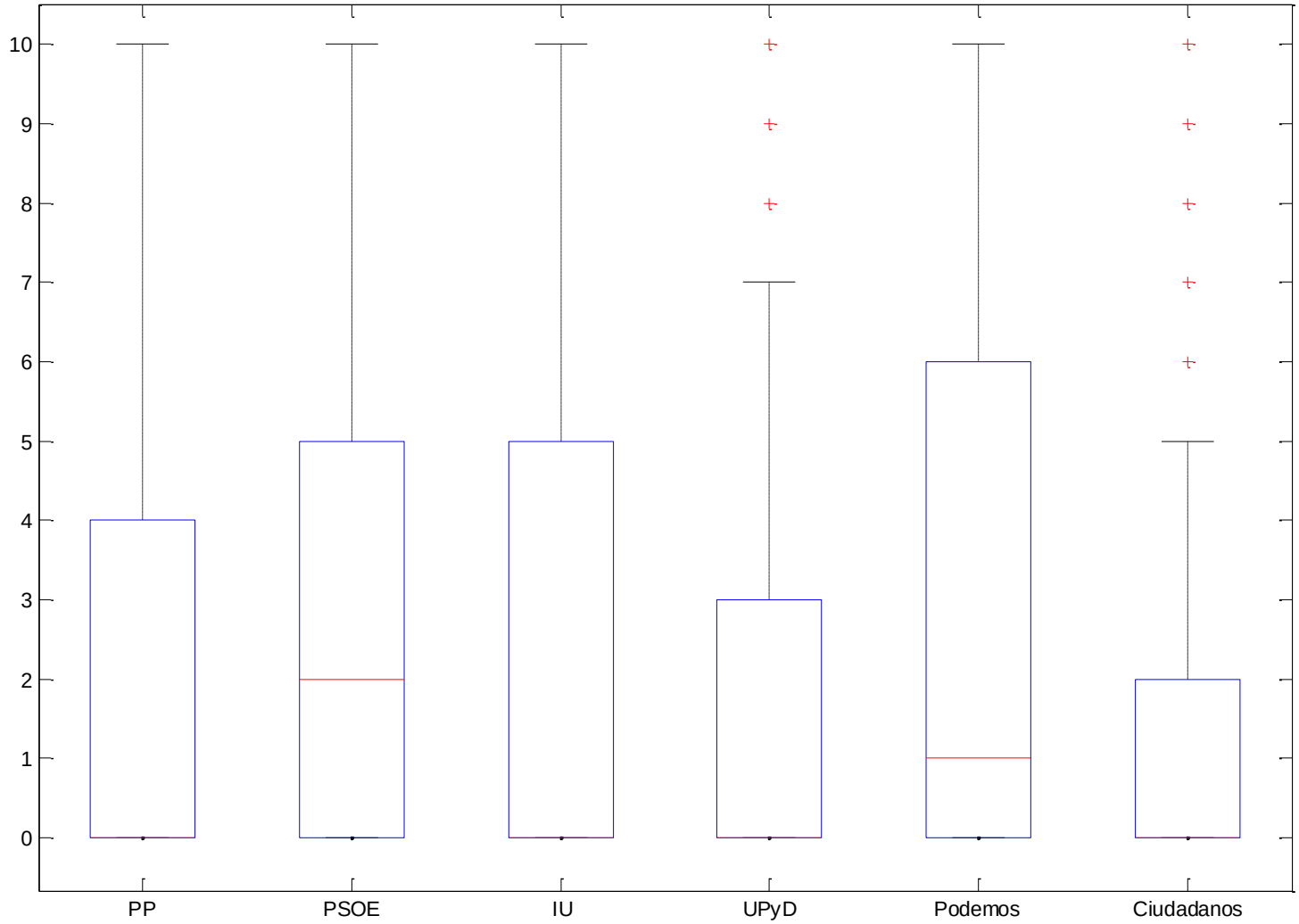
Medidas descriptivas

	Media	Desviación típica	(N)
PP	2,00	3,14	(2.305)
PSOE	2,82	3,19	(2.292)
IU (ICV en Cataluña)	2,13	2,79	(2.258)
UPyD	1,43	2,32	(2.167)
Podemos	3,06	3,50	(2.146)
Ciudadanos	1,42	2,50	(1.966)

	Media	Desviación típica	(N)
PP	8,24	1,71	(2.038)
PSOE	4,61	1,83	(2.012)
IU (ICV en Cataluña)	2,67	1,42	(1.931)
UPyD	5,55	2,00	(1.486)
Podemos	2,43	1,59	(1.611)
Ciudadanos	5,38	2,21	(1.112)

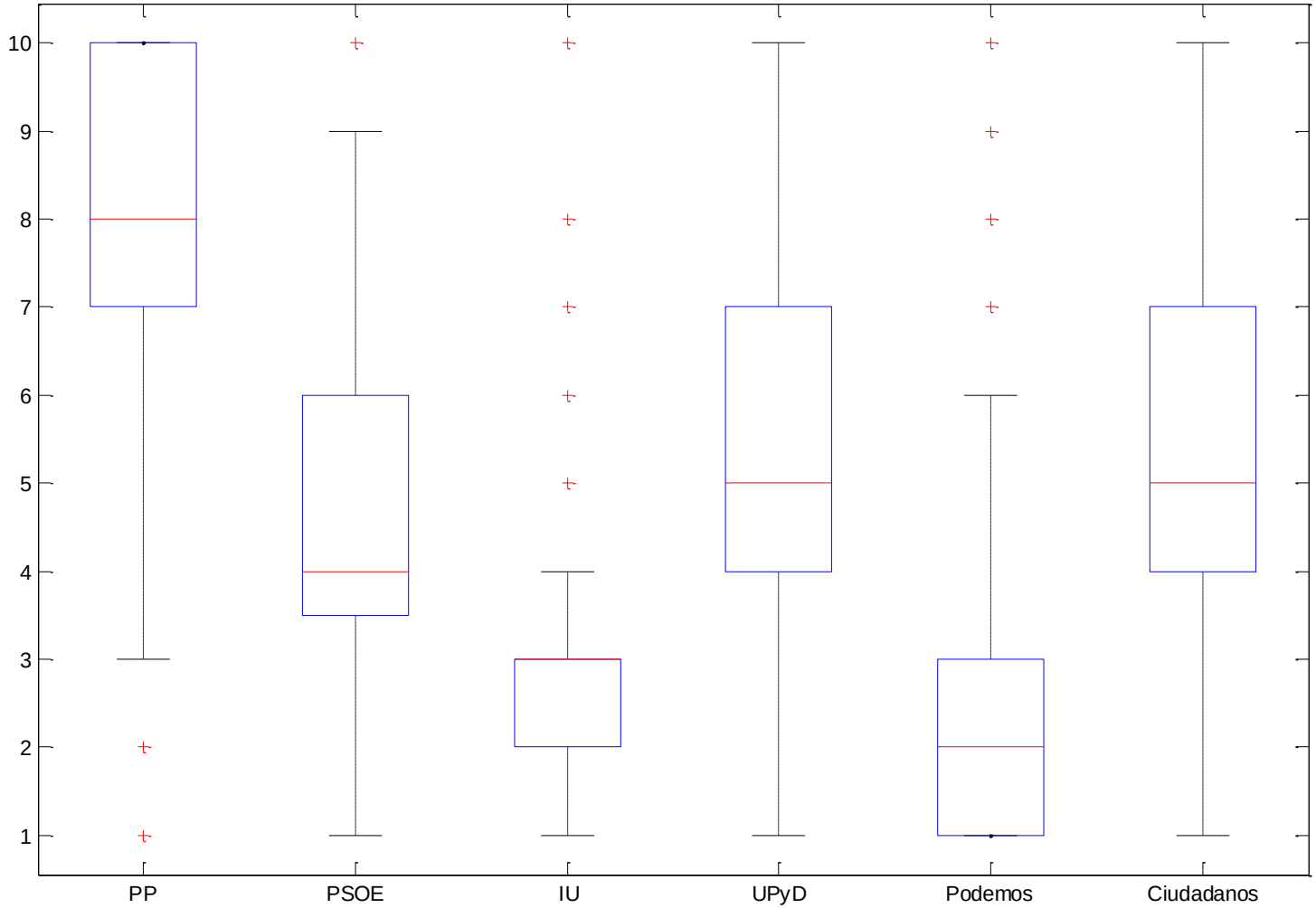
Los gráficos siguientes se basan en una reconstrucción del conjunto de datos a partir de los resultados tabulados de las preguntas 16 y 25 del Estudio 3041.

Diagrama de caja



Probabilidad (escala) de voto

Diagrama de caja



Clasificación Izquierda - Derecha

ii) Pictogramas

Veamos que dice la RAE:

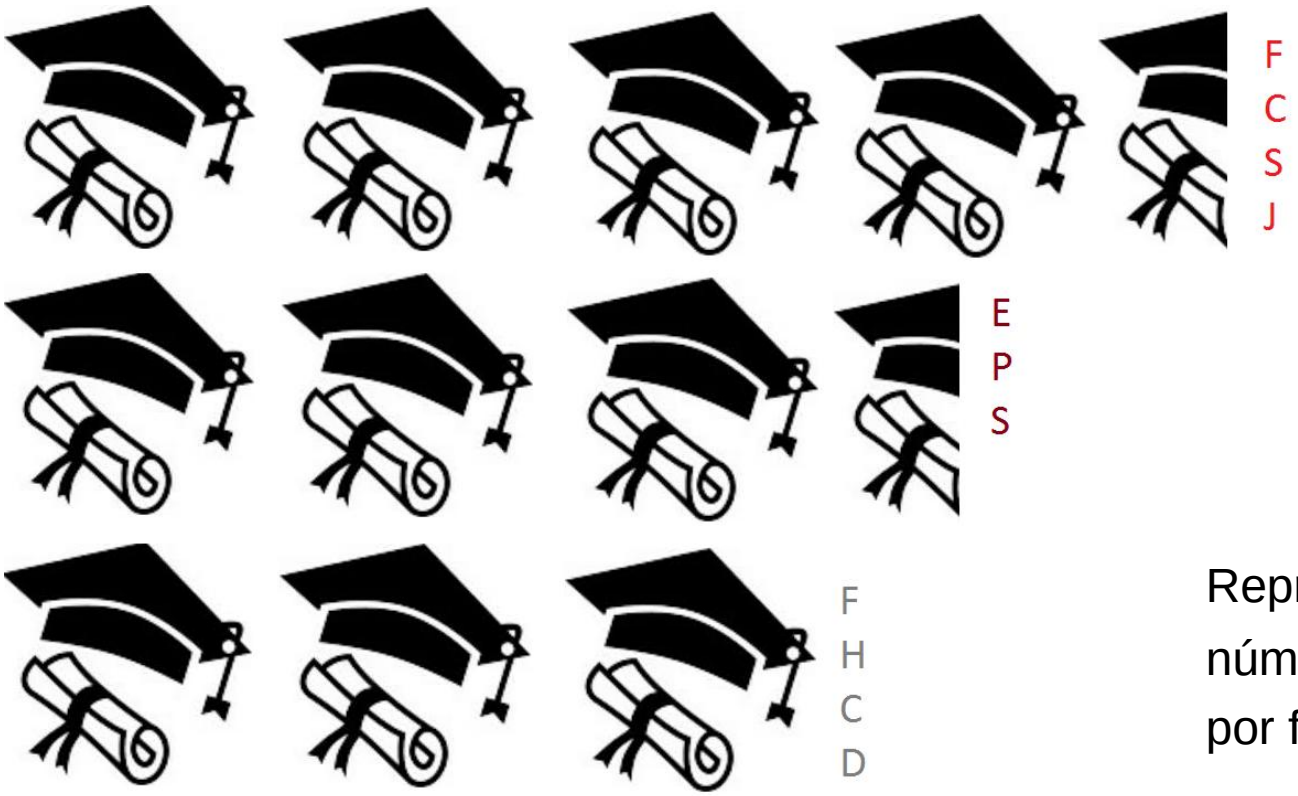
pictograma.
(Del lat. *pictus*, pintado, y *-grama*).
1. m. Signo de la escritura de figuras o símbolos.

y Wikipedia:

Es un diagrama que utiliza imágenes o símbolos para mostrar datos para una rápida comprensión. En un **pictograma**, se utilizan imágenes o símbolos para representar una cantidad específica y su tamaño o cantidad es proporcional a la frecuencia que representa.

Para realizarlo primero se escogen figuras alusiva al tema y se le asigna una imagen. En caso de que una cantidad represente un valor decimal, la figura aparece mutilada.

ii) Pictogramas



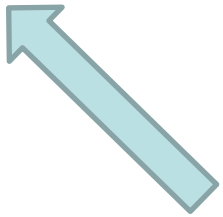
Representación del número de graduados por facultades

Faltaría una leyenda donde nos informen del significado de cada birrete.

 = 100 graduados

ii) Pictogramas

La definición nos dice que “en caso de que una cantidad represente un valor decimal, la figura aparece mutilada”, pero lo más utilizado es emplear la misma figura con un **dimensión proporcional** a la frecuencia que representa.



PELIGRO DE MAL USO

En ocasiones, el diseñador del pictograma confunde dimensión con altura y, en tal caso, el pictograma está incorrectamente escalado.

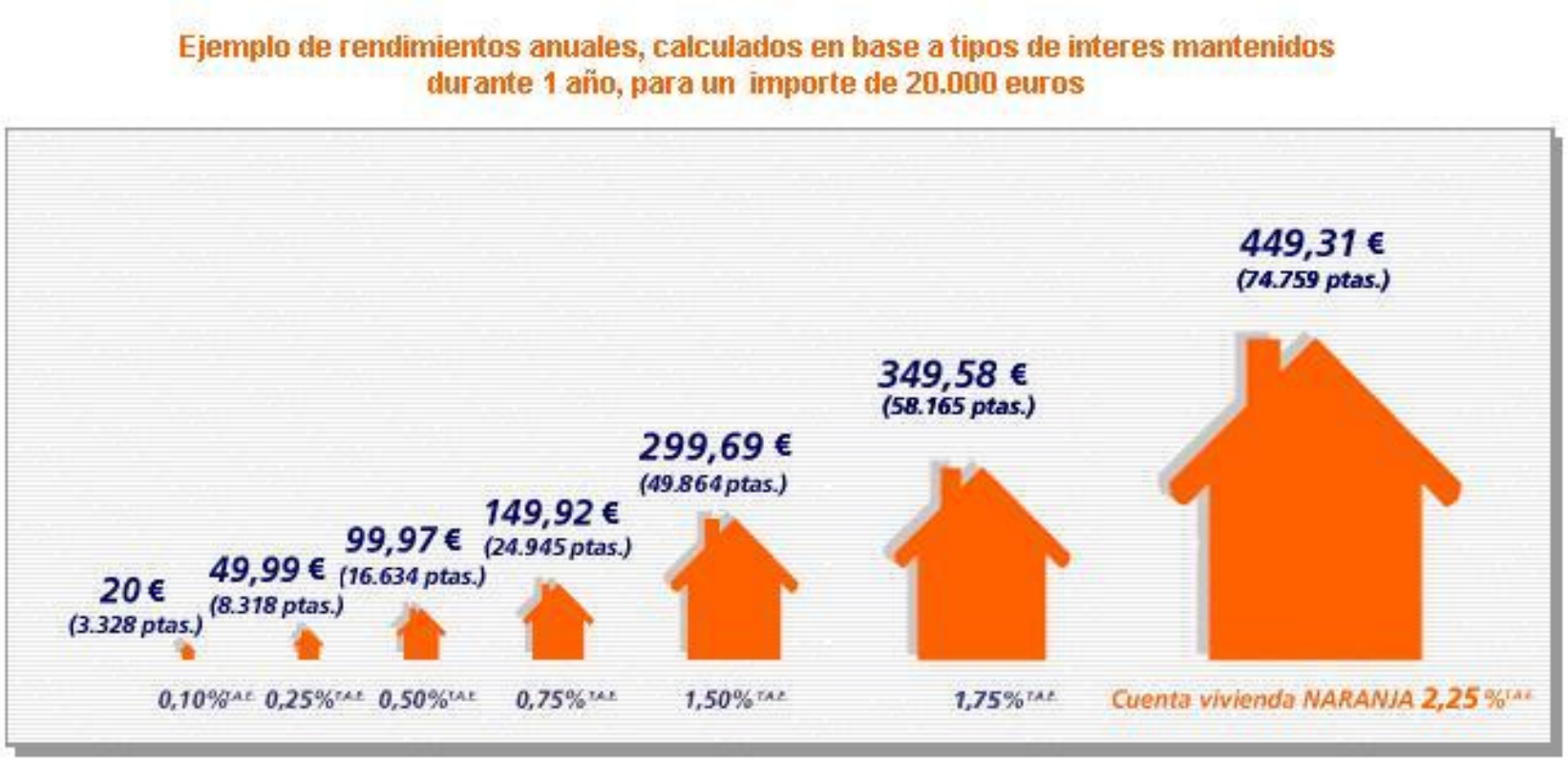
ii) Pictogramas

Si para no malinterpretar un gráfico, tengo que mirar los números con detenimiento, entonces no es un buen gráfico.

La moneda es redonda, su **dimensión** es πr^2 y aquí se ha usado incorrectamente el valor 2r o diámetro.

Si el objetivo era mostrar la tendencia creciente, bastaba con unas barras o un diagrama de líneas.

¿Es correcto?



Desafortunadamente es un error muy frecuente, incluso entre educadores:

http://www.ite.educacion.es/formacion/materiales/86/cd/m3/prctica_guiada_2.html

iii) Frecuencias absolutas, relativas y condicionadas

Veamos que dice la RAE:

> **frecuencia.**
> (Del lat. *frequentĭa*).
> **3.** f. *Estad.* Número de elementos comprendidos dentro de un intervalo en una distribución determinada.

y Wikipedia:

> La **frecuencia absoluta** es el número de veces que aparece un determinado valor en un estudio estadístico. Se representa por n_i.
>
> La **frecuencia relativa** es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento y se representa por f_i.

Cuidado con Wikipedia, hay errores, por ejemplo en la definición de <Frecuencia estadística>

iii) Frecuencias absolutas, relativas y condicionadas

Pregunta 35

¿Cómo calificaría Ud. su situación económica personal en la actualidad: es muy buena, buena, regular, mala o muy mala?

	Total	Hombre	Mujer
Muy buena	26	18	8
Buena	613	307	306
Regular	1207	561	646
Mala	456	234	222
Muy mala	176	90	86
N.S.	1	1	0
N.C.	1	0	1
(N)	2.480	1.211	1.269

Frecuencias absolutas

A partir de esta tabla, podemos decir "*325 hombres y 314 mujeres opina que su situación económica es buena o muy buena*" … **¿son muchos?**

Esta tabla es una reconstrucción a partir de la tabla oficial que aparece más adelante.

iii) Frecuencias absolutas, relativas y condicionadas

Pregunta 35

¿Cómo calificaría Ud. su situación económica personal en la actualidad: es muy buena, buena, regular, mala o muy mala?

	Total	Hombre	Mujer
Muy buena	1,0	0,7	0,3
Buena	24,7	12,4	12,3
Regular	48,7	22,6	26,0
Mala	18,4	9,4	9,0
Muy mala	7,1	3,6	3,5
N.S.	0,0	0,0	0,0
N.C.	0,0	0,0	0,0
(N)	100,0	48,8	51,2

Frecuencias relativas

Con estas frecuencias podríamos **comparar** los resultados obtenidos en estudios similares sin que los tamaños de los estudios tengan que ser iguales o parecidos.

iii) Frecuencias absolutas, relativas y condicionadas

	Total	Hombre	Mujer
Muy buena	1,0	0,7	0,3
Buena	24,7	12,4	12,3
Regular	48,7	22,6	26,0
Mala	18,4	9,4	9,0
Muy mala	7,1	3,6	3,5
N.S.	0,0	0,0	0,0
N.C.	0,0	0,0	0,0
(N)	100,0	48,8	51,2

Frecuencias relativas

"*El 13,1% de los hombres y el 12,6% de las mujeres y opinan que su situación económica es buena o muy buena*".

iii) Frecuencias absolutas, relativas y condicionadas

	Total	Hombre	Mujer
Muy buena	1,0	0,7	0,3
Buena	24,7	12,4	12,3
Regular	48,7	22,6	26,0
Mala	18,4	9,4	9,0
Muy mala	7,1	3,6	3,5
N.S.	0,0	0,0	0,0
N.C.	0,0	0,0	0,0
(N)	100,0	48,8	51,2

Frecuencias relativas

"*El 13,1% de los hombres y el 12,6% de las mujeres y opinan que su situación económica es buena o muy buena*".

"*El 13,1%* (12,6%) *de los encuestados son hombres* (mujeres) *y opinan que su situación económica es buena o muy buena*".

iii) Frecuencias absolutas, relativas y condicionadas

	TOTAL	Sexo de la persona entrevistada	
		Hombre	Mujer
Muy buena	1,0	1,5	0,6
Buena	24,7	25,4	24,1
Regular	48,7	46,3	50,9
Mala	18,4	19,3	17,5
Muy mala	7,1	7,4	6,8
N.S.	0,0	0,1	-
N.C.	0,0	-	0,1
(N)	(2.480)	(1.211)	(1.269)

Las tablas publicadas por agencias como el CIS suelen reportar **frecuencias condicionadas**, es decir, la frecuencia relativa se calcula sobre el total de individuos que cumplen una condición.

En el ejemplo: Se divide por los totales por las columnas.

"*El 26,9% de los hombres y el 24,7% de las mujeres opinan que su situación económica es buena o muy buena*".

	TOTAL	Sexo de la persona entrevistada	
		Hombre	Mujer
Muy buena	1,0	1,5	0,6
Buena	24,7	25,4	24,1
Regular	48,7	46,3	50,9
Mala	18,4	19,3	17,5
Muy mala	7,1	7,4	6,8
N.S.	0,0	0,1	-
N.C.	0,0	-	0,1
(N)	(2.480)	(1.211)	(1.269)

"*El 2,1% de los encuestados opinan que su situación económica es muy buena*".

"*Entre los encuestados con condiciones económicas muy malas, el 7,4% son hombres*".

	TOTAL	Sexo de la persona entrevistada	
		Hombre	Mujer
Muy buena	1,0	1,5	0,6
Buena	24,7	25,4	24,1
Regular	48,7	46,3	50,9
Mala	18,4	19,3	17,5
Muy mala	7,1	7,4	6,8
N.S.	0,0	0,1	-
N.C.	0,0	-	0,1
(N)	(2.480)	(1.211)	(1.269)

"*El 2,1% de los encuestados opinan que su situación económica es muy buena*".

"*El 1,0% de los encuestados opinan que su situación económica es muy buena*".

Es incorrecto sumar frecuencias condicionadas cuando se condiciona en valores distintos.

	TOTAL	Sexo de la persona entrevistada	
		Hombre	Mujer
Muy buena	1,0	1,5	0,6
Buena	24,7	25,4	24,1
Regular	48,7	46,3	50,9
Mala	18,4	19,3	17,5
Muy mala	7,1	7,4	6,8
N.S.	0,0	0,1	-
N.C.	0,0	-	0,1
(N)	(2.480)	(1.211)	(1.269)

"*Entre los encuestados con condiciones económicas muy malas, el 7,4% son hombres*".

"*Entre los encuestados con condiciones económicas muy malas, el 51,1% son hombres*".

Notar que esta tabla es condicional por columnas (sexo del entrevistado) y la afirmación se refiere a una condición por filas (situación económica personal).

iv) Intervalos de confianza

Veamos que dice la RAE:

intervalo.
(Del lat. *intervallum*).
2. m. Conjunto de los valores que toma una magnitud entre dos límites dados. *Intervalo de temperaturas, de energías, de frecuencias.*

confianza.
1. f. Esperanza firme que se tiene de alguien o algo.
de ~.
3. loc. adj. Dicho de una cosa: Que posee las cualidades recomendables para el fin a que se destina.

iv) Intervalos de confianza

y Wikipedia:

> En estadística, se llama a un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido con una determinada probabilidad de acierto.
>
> Formalmente, estos números determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un *parámetro poblacional*.
>
> La probabilidad de éxito en la estimación se representa con $1 - \alpha$ y se denomina *nivel de confianza*.

Conclusión: Un intervalo de confianza se destina para estimar un parámetro poblacional desconocido.

iv) Intervalos de confianza

Corrigiendo a Wikipedia:

En estadística, se llama a un par o varios pares de ~~números~~ cantidades entre los cuales ~~se estima que~~ estará cierto valor desconocido con una determinada probabilidad de acierto.

Formalmente, estas ~~números~~ cantidades determinan un intervalo, que se calcula a partir de datos de una muestra, y el valor desconocido es un *parámetro poblacional*.

La probabilidad de éxito en la estimación se representa con $1 - \alpha$ y se denomina *nivel de confianza*.

Formalmente, los números no son aleatorios pero si pueden serlo las cantidades. Aunque en su definición se dice que estos “números” se calculan a partir de la muestra y en ese sentido su definición es correcta.

iv) Intervalos de confianza

Algunos ejemplos (con fórmulas):

Intervalo de confianza del 95% para la media:

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1}(0,975)$$

Intervalo de confianza del 95% para una proporción:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \right)$$

iv) Intervalos de confianza

Intervalo de confianza para la media (con Excel)

	A	B	C	D
1	n	150		Datos
2	media	1,77		
3	s	2,27		
4				
5	alfa	0,05		
6				
7	t*s/raíz(n)	0,36		=INTERVALO.CONFIANZA(alfa;s;n)
8				
9	intervalo	1,41	2,13	
10		=B2-B7	=B2+B7	
11				
12	Datos correspondientes a Carlos Salvador			
13	Tomado de http://www.cis.es/cis/opencms/ES/11_barometros/avances.html			

iv) Intervalos de confianza

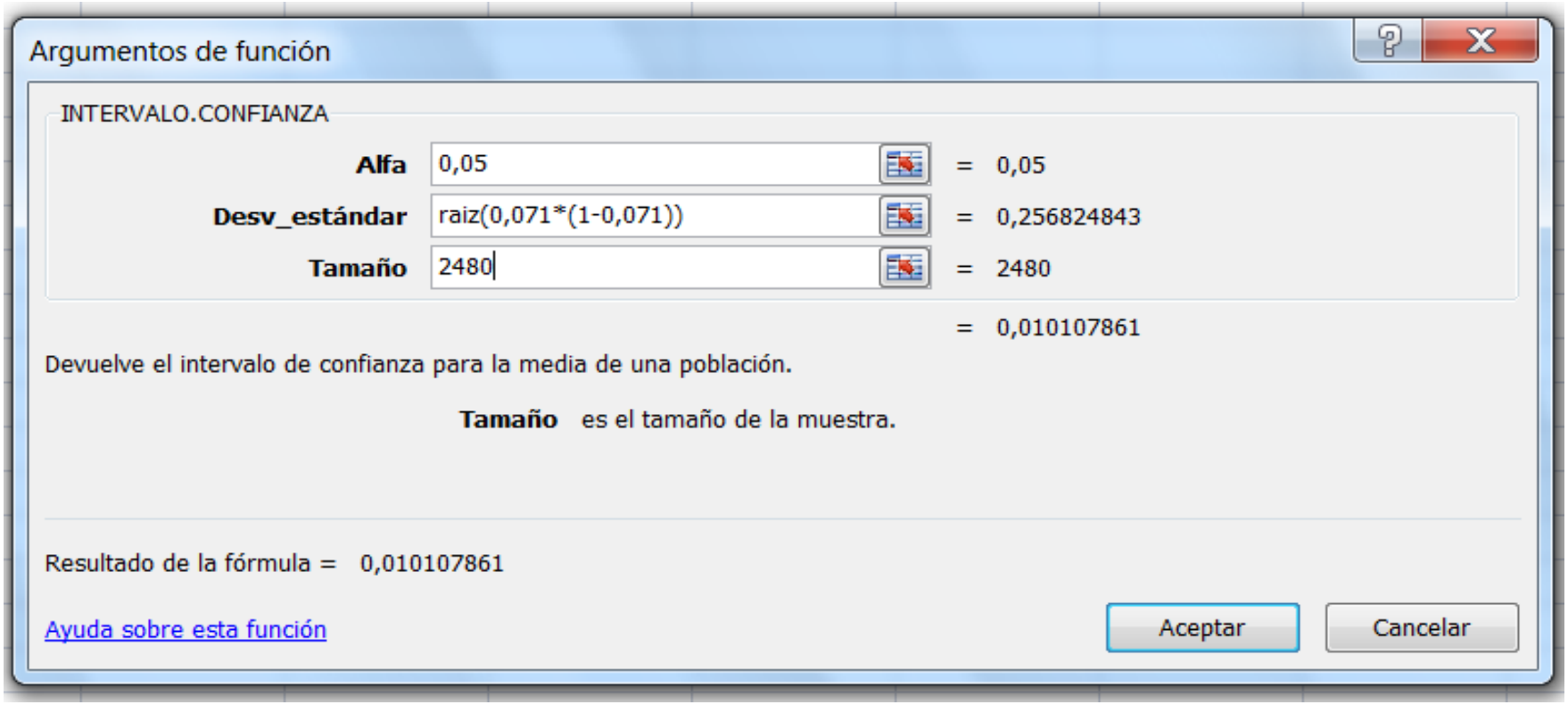
Intervalo de confianza para una proporción (con Excel)

	A	B	C	D
1	n	2480		Datos
2	x	176		
3				
4	p	0,071	=B2/B1	Proporción
5	p(1-p)/n	2,6585E-05	=B4*(1-B4)/B1	Varianza
6				
7	alfa	0,05		Computación de z
8	alfa/2	0,025		
9	1-alfa/2	0,975		
10	z	1,960		=DISTR.NORM.ESTAND.INV(0,975)
11				
12	z*raíz(p(1-p)/n)	0,010		
13				
14	Intervalo	0,061	0,081	
15		=B4-B12	=B4+B12	
16				
17	Datos correspondientes a categoría "Muy mala" en pregunta 35.			

iv) Intervalos de confianza

¿Podemos utilizar INTERVALO.CONFIANZA para una proporción?

¡Si! … pero tiene "truco"



Ahora la desviación estándar es $\sqrt{}$ (p x (1-p)).

iv) Intervalos de confianza

Intensión de voto …

En este ejemplo cambiaremos al estudio 3033 de Julio/2014 puesto que es “equivalente” en el cuestionario al estudio 3041 de Octubre/2014 y están disponibles los microdatos.

Totalmente de acuerdo con esta aclaración que aparece en los estudios del CIS.

> Dado que los datos de los indicadores “intención de voto” e “intención de voto + simpatía” son datos directos de opinión y no suponen ni proporcionan por sí mismos ninguna proyección de hipotéticos resultados electorales, en este anexo se recogen los resultados de aplicar un modelo de estimación a los datos directos de opinión proporcionados por la encuesta. Procedimiento que conlleva la ponderación de los datos por recuerdo de voto imputado y aplicación de modelos que relacionan la intención de voto con otras variables. Obviamente, la aplicación a los mismos datos de otros modelos podría dar lugar a estimaciones diferentes.

iv) Intervalos de confianza

	A	B	C	D	E	F	G	H
1		Estimación KNN del voto		Horquilla (Lim Inf - Lim Sup)				
2	*PP*	22,5%		20,6%	24,5%		1,9%	n = 1766
3	*PSOE*	20,5%		18,6%	22,4%		1,9%	
4	*IU (ICV en Cataluña)*	10,0%		8,6%	11,4%		1,4%	
5	*UPyD*	5,9%		4,8%	7,0%		1,1%	
16	*Podemos*	18,3%		16,5%	20,1%		1,8%	
20								
21		Estimación del voto CIS		Horquilla (Lim Inf - Lim Sup)				
22	*PP*	30,0%		28,0%	32,0%		2,0%	n = 1992
23	*PSOE*	21,2%		19,4%	23,0%		1,8%	
24	*IU (ICV en Cataluña)*	8,2%		7,0%	9,4%		1,2%	
25	*UPyD*	5,9%		4,9%	6,9%		1,0%	
36	*Podemos*	15,3%		13,7%	16,9%		1,6%	

Modelos diferentes pueden conducir a “predicciones” diferentes

iv) Intervalos de confianza

	A	B	C	D	E	F	G	H
1		Estimación KNN del voto		Horquilla (Lim Inf - Lim Sup)				
2	*PP*	22,5%		20,6%	24,5%		1,9%	n = 1766
3	*PSOE*	20,5%		18,6%	22,4%		1,9%	
4	*IU (ICV en Cataluña)*	10,0%		8,6%	11,4%		1,4%	
5	*UPyD*	5,9%		4,8%	7,0%		1,1%	
16	*Podemos*	18,3%		16,5%	20,1%		1,8%	
20								

¿En qué es mejor este modelo KNN al del CIS?

Al menos en una cosa:

Es transparente en el cómo se obtiene la estimación del voto.

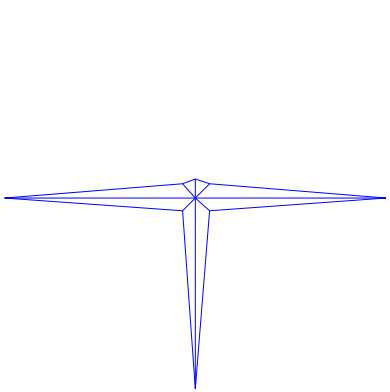
iv) Intervalos de confianza

Un último apunte sobre intervalos:

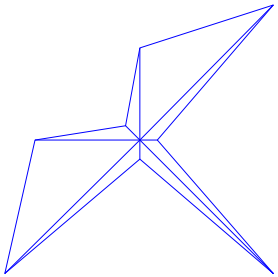
	A	B	C	D	E	F	G
1		Estimación KNN del voto		Horquilla (Lim Inf - Lim Sup)			
2	*IU (ICV en Cataluña)*	10,0%		8,6%	11,4%		1,4%
3	*Podemos*	18,3%		16,5%	20,1%		1,8%
4	*IU + Podemos*	28,3%		26,2%	30,4%		2,1%
5							
6							
7	*UPyD*	5,9%		4,8%	7,0%		1,1%
8	*Ciudadanos*	1,0%		0,6%	1,5%		0,5%
9	*UPyD + Ciudadanos*	6,9%		5,7%	8,1%		1,2%
10							

Los “escaños” se suman pero los intervalos NO.

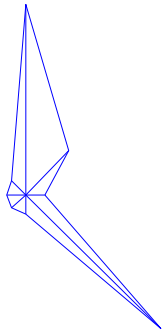
Perfiles de votantes



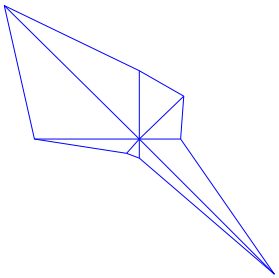
PP



PSOE



IU (ICV en Cataluña)



UPyD

Podemos

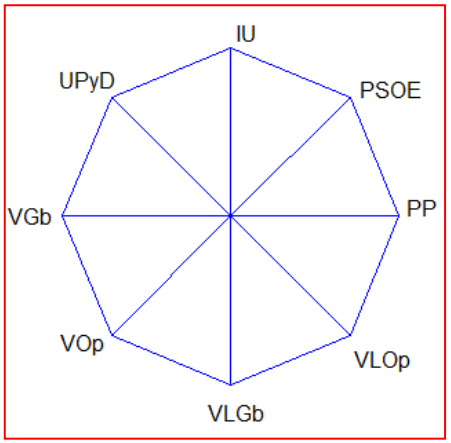
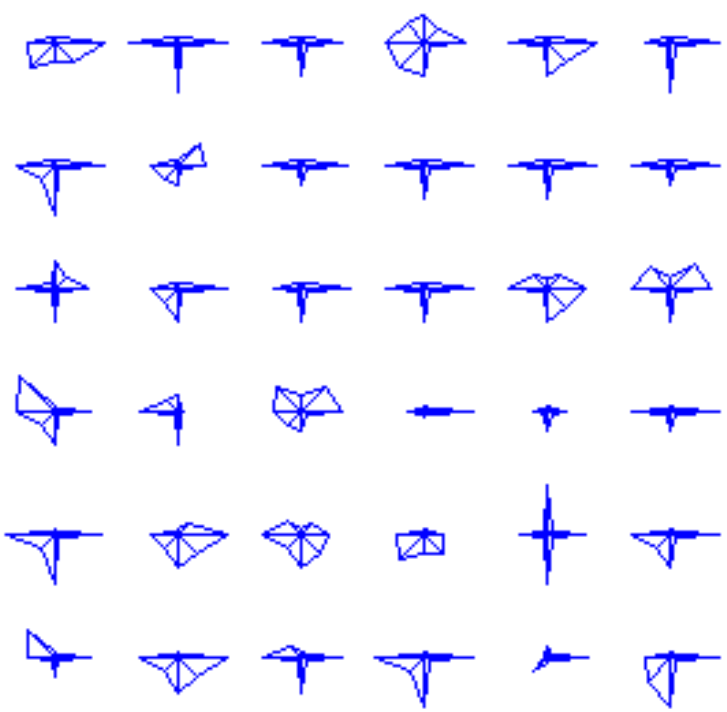
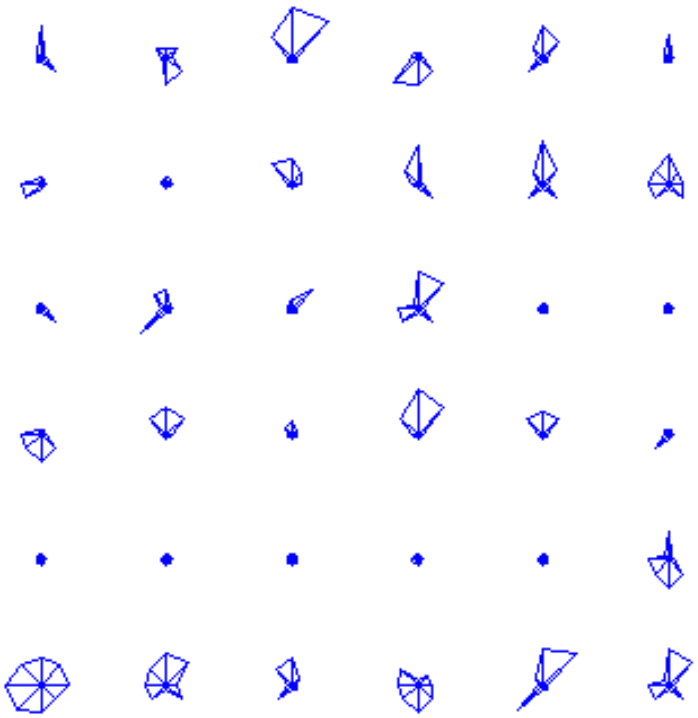


Gráfico basado en los valores medianos

¿Todos los votantes son “iguales”?

Muestras de perfiles entre votantes directos



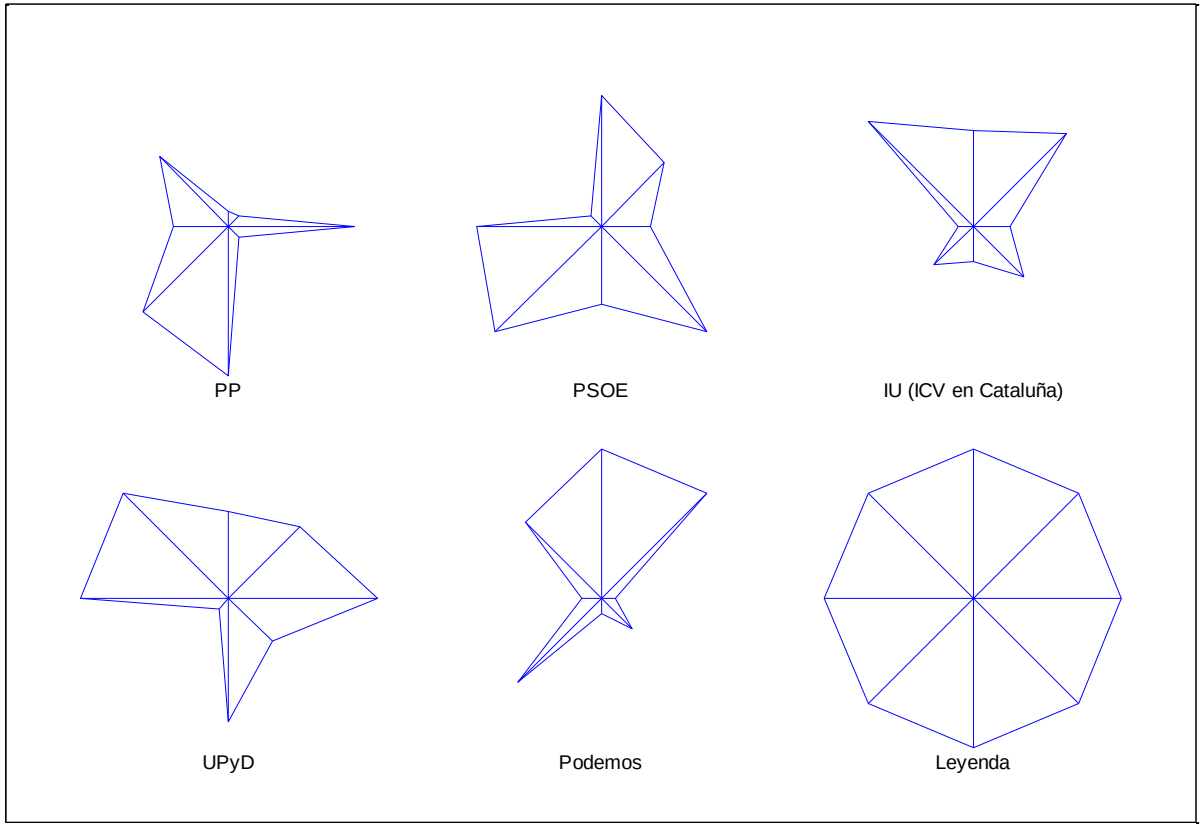


Por supuesto que no …

¿Todos los votantes son “iguales”? No

¿Cómo evaluar la “diversidad” de los votantes?

Podemos usar el mismo tipo de gráficos pero con medidas de dispersión.



¿Todos los profesores de estadística son iguales?

No lo sé

Pero contad conmigo para analizar
los datos de esa encuesta …