Chapter 2: Analysis of univariate data

Objective:

Show how graphics and numerical measures can be used to summarise the main features of a data set.

Outline:

- Frequency tables.
- Graphical methods for qualitative data: pie and bar charts, …
- Graphical methods for discrete data: bar charts.
- Graphical methods for continuous data: histograms …
- Numerical summaries
  - Measures of location: mode, median, mean, …
  - Measures of spread: range, iqr, standard deviation, …
  - Measures of form: skewness, kurtosis, …

Recommended reading:

- A nice video on histograms and frequency polygons

Description of qualitative variables

SAMPLE: 70 madrileño university students
VARIABLE: Preferred political party

PP	IU	Others	PP	PSOE	Others	Others
IU	PP	IU	PSOE	PSOE	UPD	IU
PP	PSOE	IU	PP	PSOE	Others	PSOE
IU	IU	PSOE	IU	IU	PSOE	PSOE
PP	PSOE	PP	PP	PSOE	IU	UPD
PP	PSOE	UPD	PSOE	PP	Others	IU
IU	PSOE	IU	PP	PSOE	IU	PSOE
IU	IU	PSOE	UPD	UPD	IU	PP
PSOE	IU	PSOE	IU	PP	PSOE	IU
PSOE	PSOE	UPD	UPD	PP	PP	PSOE

The frequency table

Absolute frequency

Relative frequency

Class (i)	n_i	f_i
PSOE	23	0,33
PP	15	0,21
IU	20	0,29
UPD	7	0,10
Others	5	0,07
Total	70	1

= 15/70

What is the modal class?

= 23+15+20+7+5

= 0,33+0,21+ …+0,07

The general outline of a frequency table

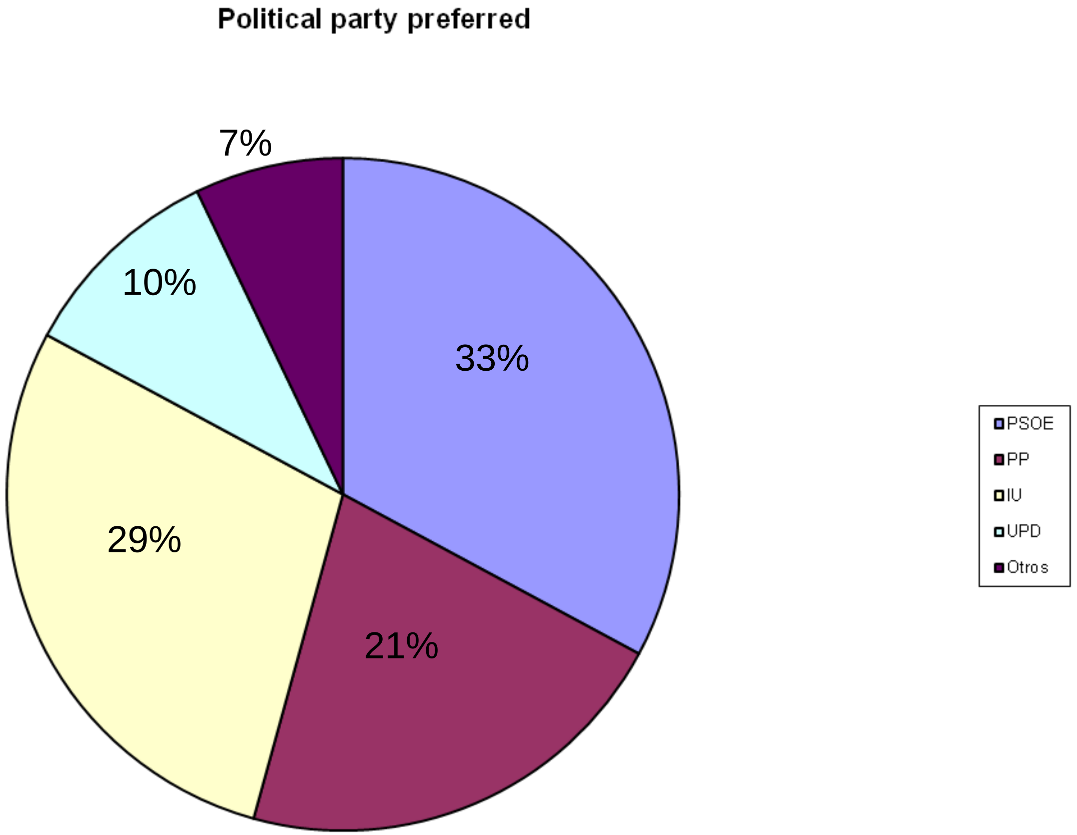
Class (i)	n_i	f_i
1	n_1	f_1
2	n_2	f_2
3	n_3	f_3
⋮	⋮	⋮
k	n_k	f_k
Total	N	1

$f_1 = n_1/N$

$N = n_1 + n_2 + \dots + n_k$

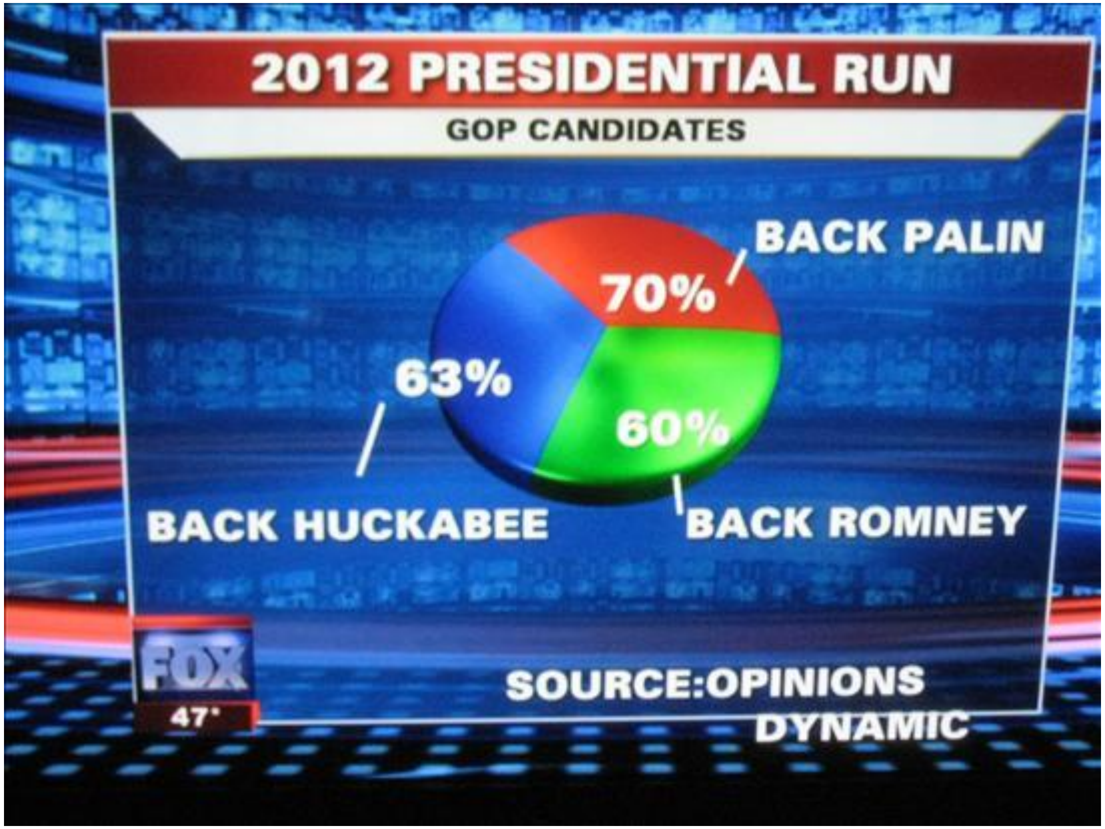
$1 = f_1 + f_2 + \dots + f_k$

The pie chart



Could we use a pie chart for other types of data?

Dodgy pie charts I



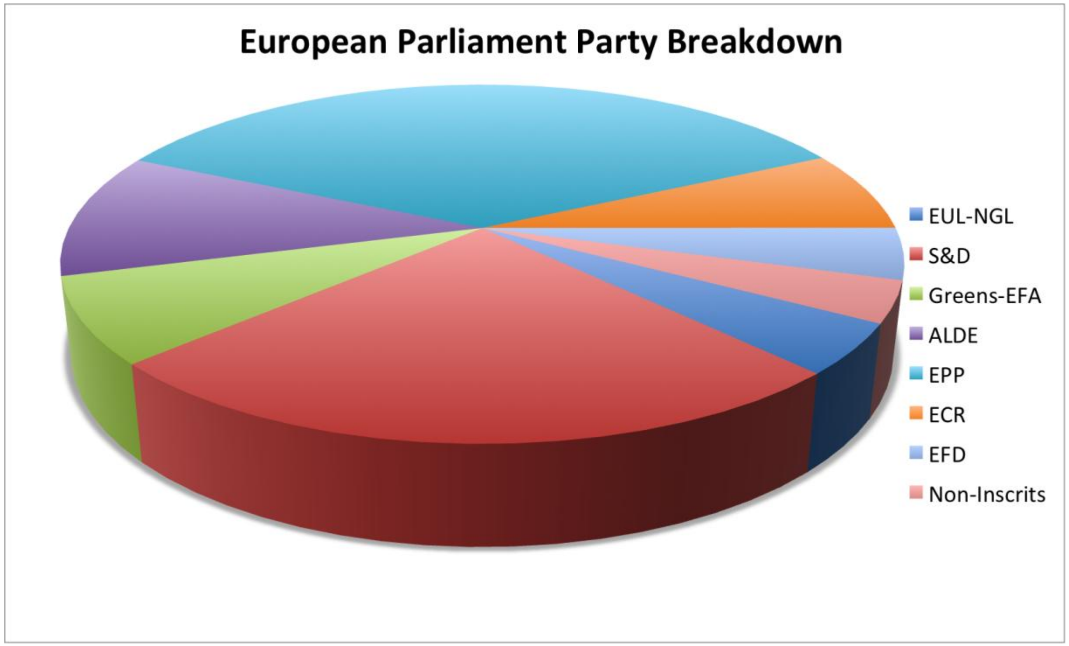
The chart shows preferences for different US candidates.

Any comments?

Any explanation?

Flowing data

Dodgy pie charts II

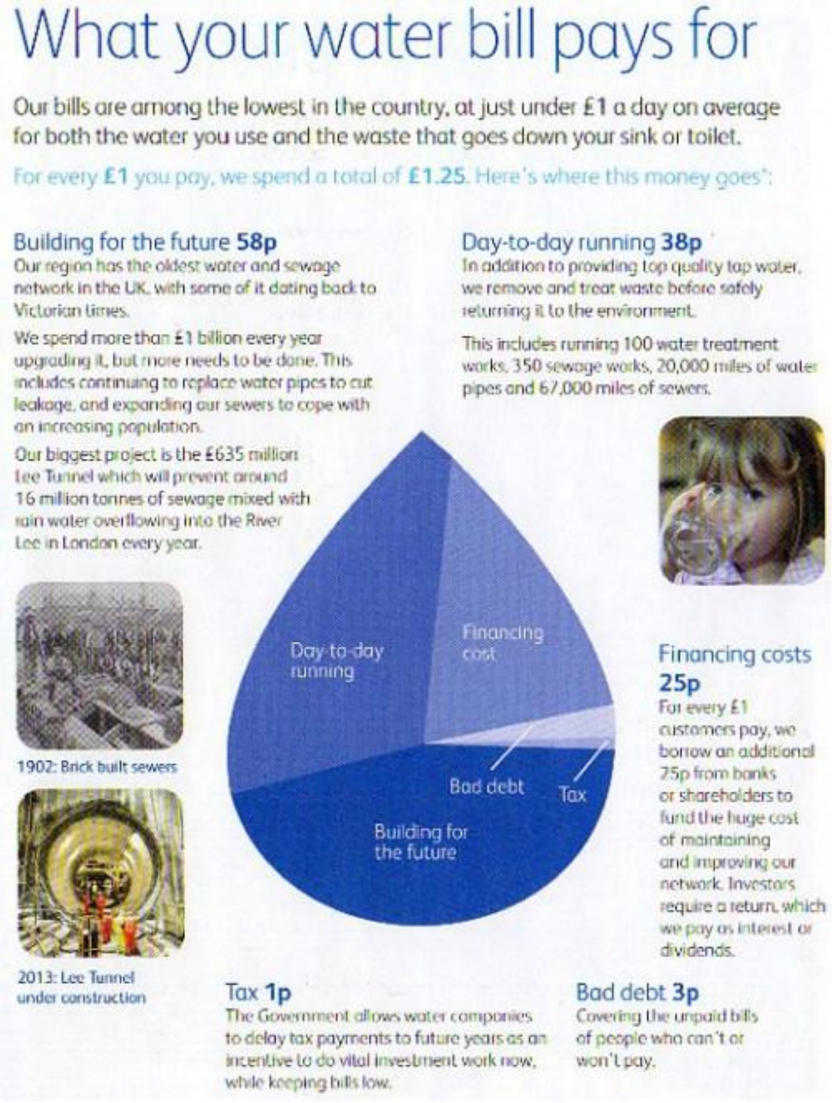


Are 3d pie charts a good idea?

Business insider

Dodgy pie charts III

The idea is to make the image more attractive, but ...



What your water bill pays for

Our bills are among the lowest in the country, at just under £1 a day on average for both the water you use and the waste that goes down your sink or toilet.

For every £1 you pay, we spend a total of £1.25. Here's where this money goes*:

Building for the future 58p

Our region has the oldest water and sewage network in the UK, with some of it dating back to Victorian times.

We spend more than £1 billion every year upgrading it, but more needs to be done. This includes continuing to replace water pipes to cut leakage, and expanding our sewers to cope with an increasing population.

Our biggest project is the £635 million Lee Tunnel which will prevent around 16 million tonnes of sewage mixed with rain water overflowing into the River Lee in London every year.

1902: Brick built sewers

2013: Lee Tunnel under construction

Day-to-day running 38p

In addition to providing top quality tap water, we remove and treat waste before safely returning it to the environment.

This includes running 100 water treatment works, 350 sewage works, 20,000 miles of water pipes and 67,000 miles of sewers.

Financing costs 25p

For every £1 customers pay, we borrow an additional 25p from banks or shareholders to fund the huge cost of maintaining and improving our network. Investors require a return, which we pay as interest or dividends.

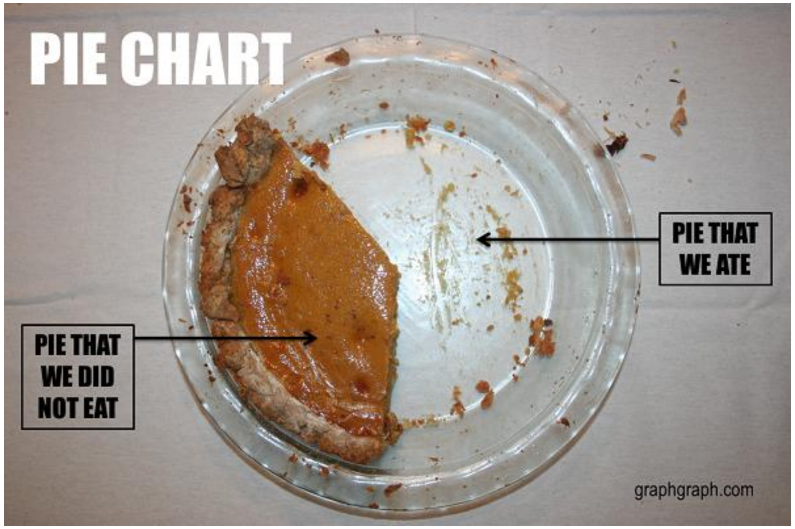
Tax 1p

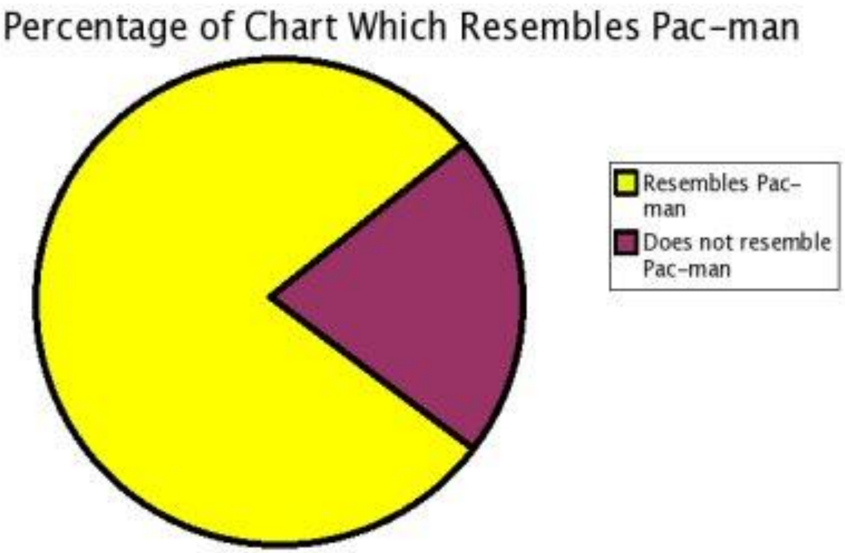
The Government allows water companies to delay tax payments to future years as an incentive to do vital investment work now, while keeping bills low.

Bad debt 3p

Covering the unpaid bills of people who can't or won't pay.

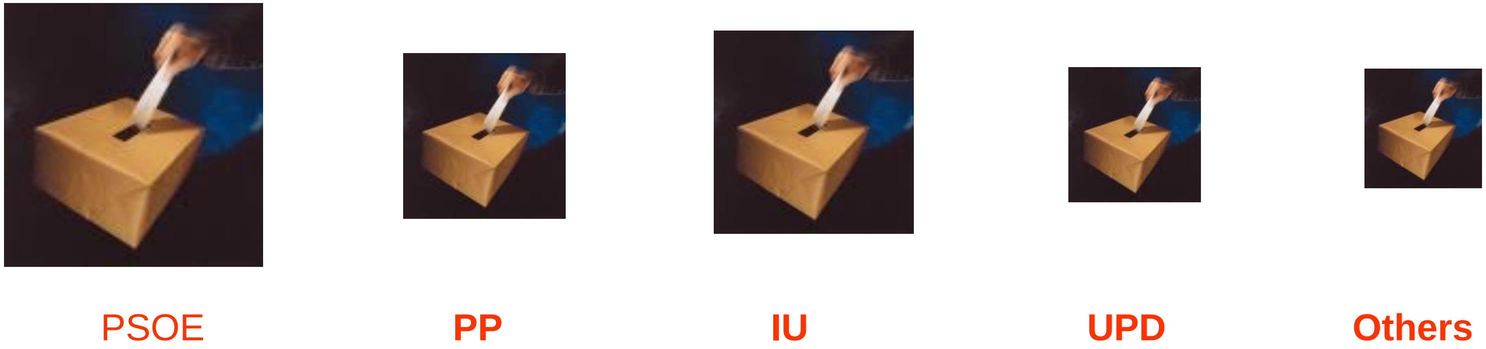
Nice pie charts





This link gives lots of other criticisms of pie charts

The pictogram

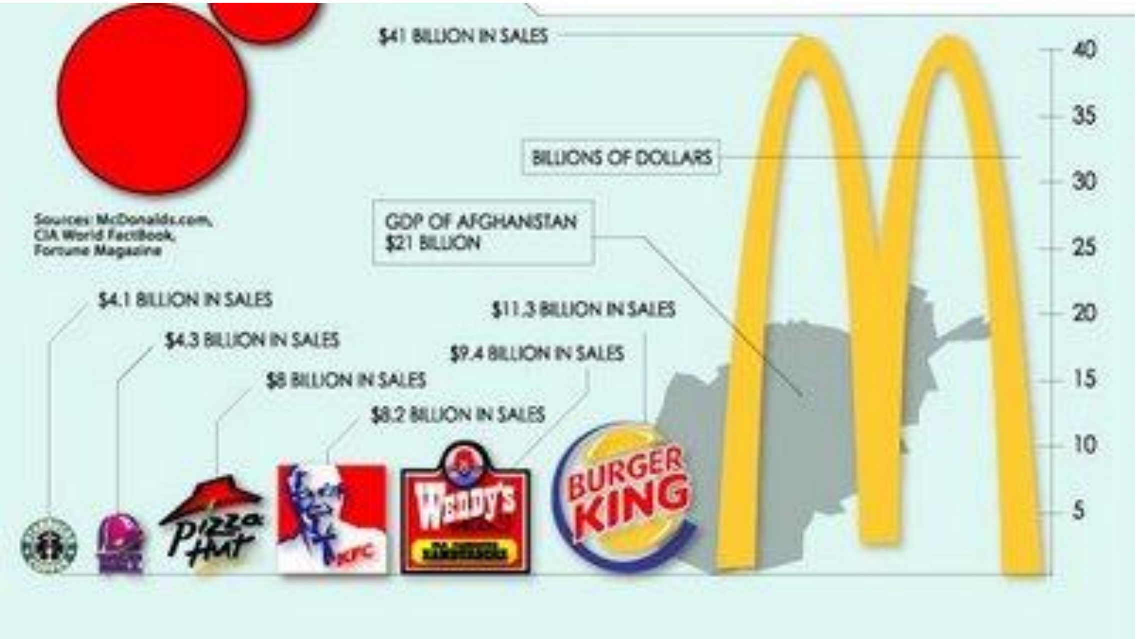


The area of the graph is proportional to the frequency.

What sort of data is this appropriate for?

What are the advantages / disadvantages compared to pie charts?

How to *lie* with pictograms

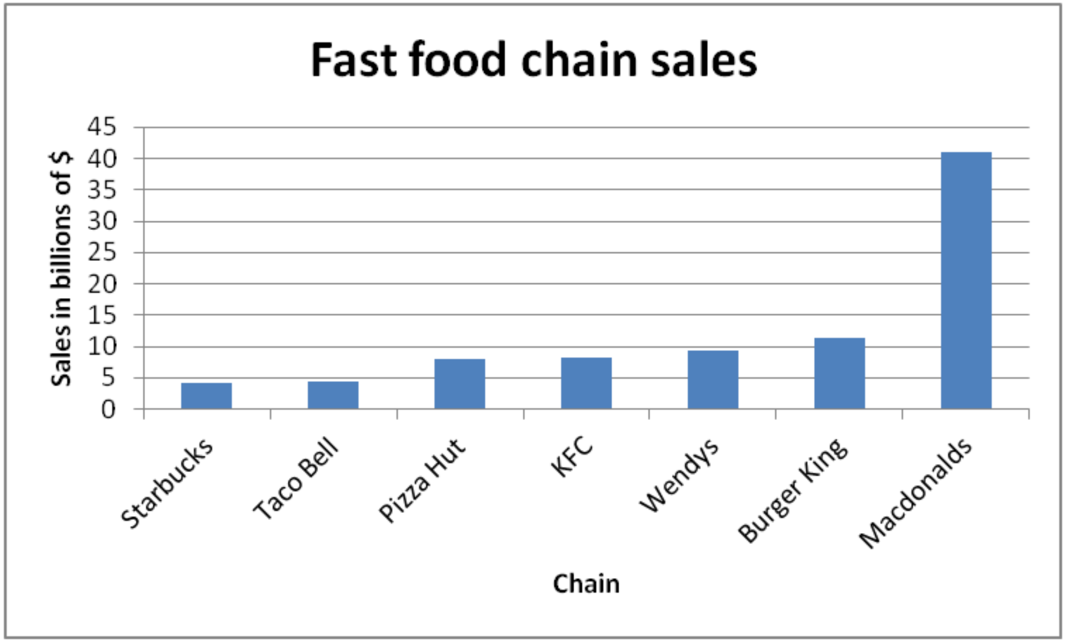


What is your impression about fast food sales?

Are there any better graphs?

Agoraphilia

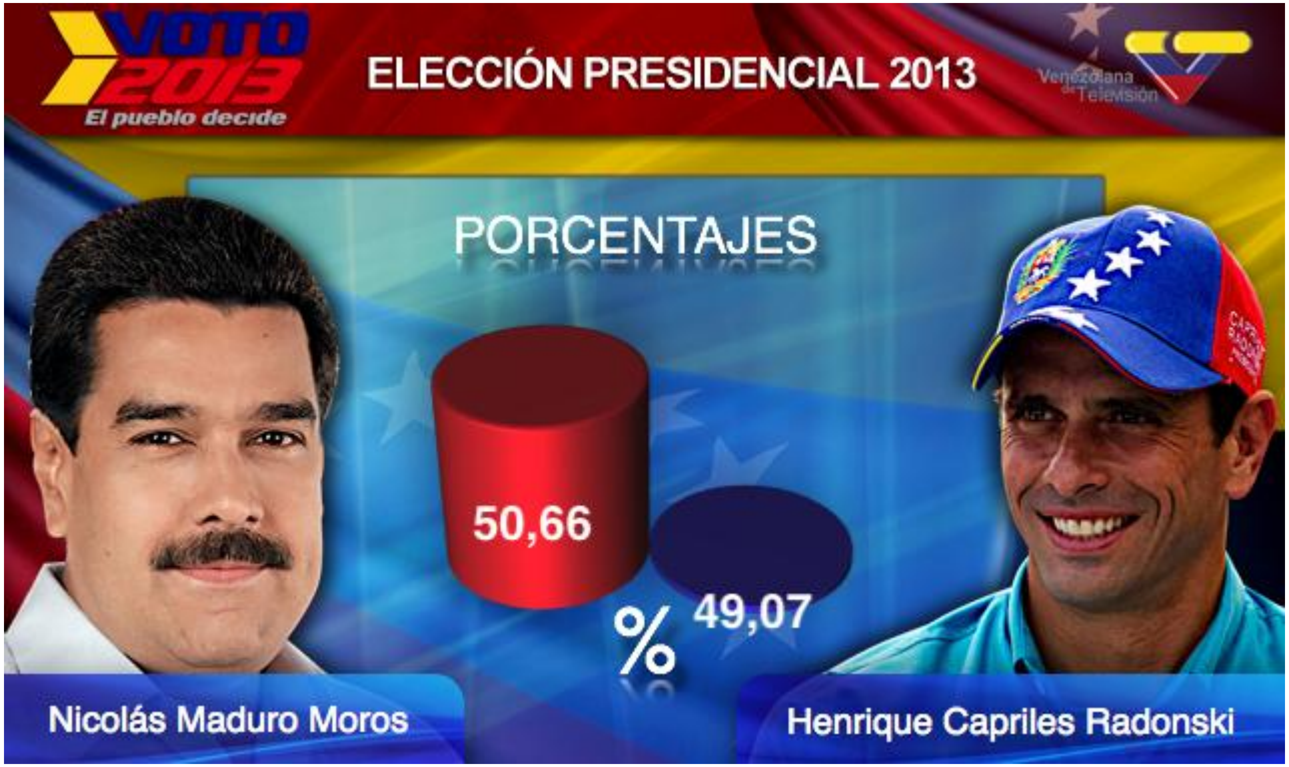
The bar chart



Will this work with other types of data?

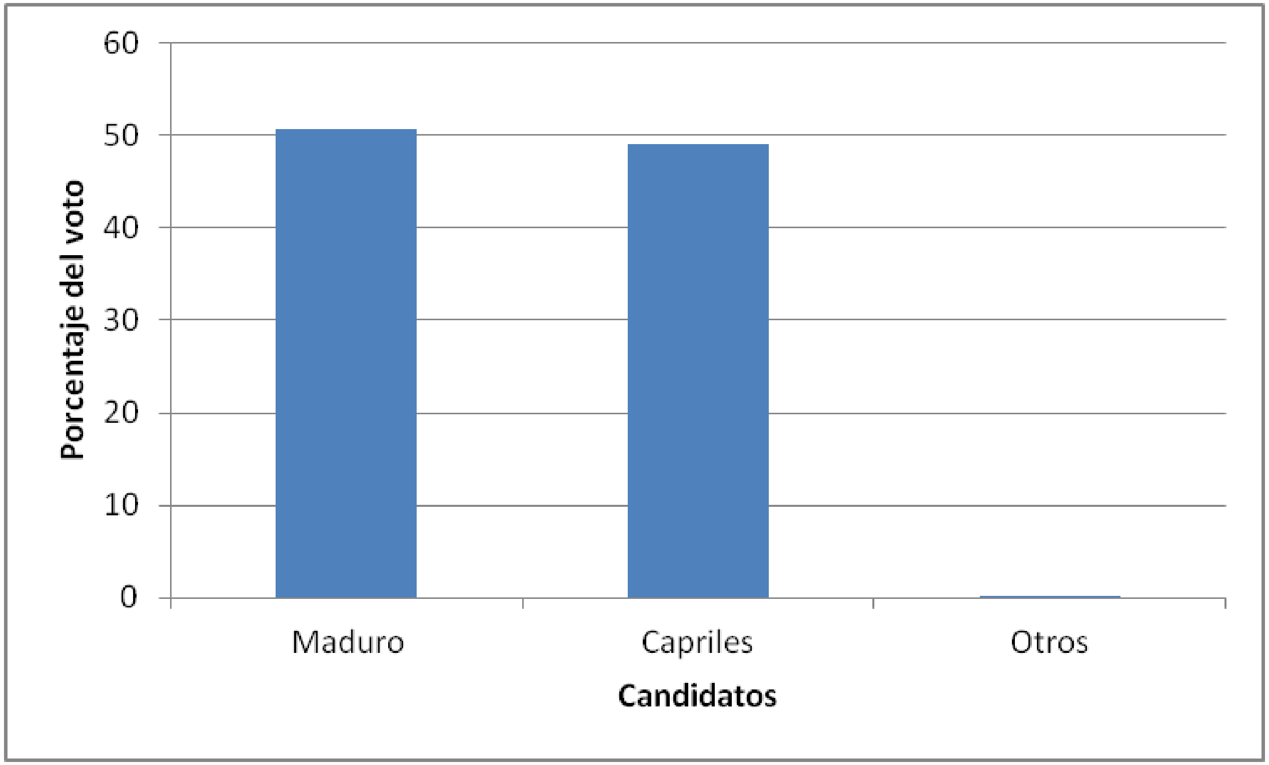
How to lie with a bar chart

The following graphic appeared on Venezuelan state tv after the 2013 elections.



It looks visually like Nicolás Maduro romped home…

… if you don’t look at the percentages!



In the previous graphic, the vertical axis has been cut to (deliberately?) give a misleading impression.

Bar charts for discrete data

The table shows the number of times that people have voted in the Community elections for a sample of 60 Madrileños.

What is the mode?

Number of times voted	Absolute frequency
0	4
1	10
2	12
3	15
4	11
5	5
6	1
7	1
8	1
Total	60

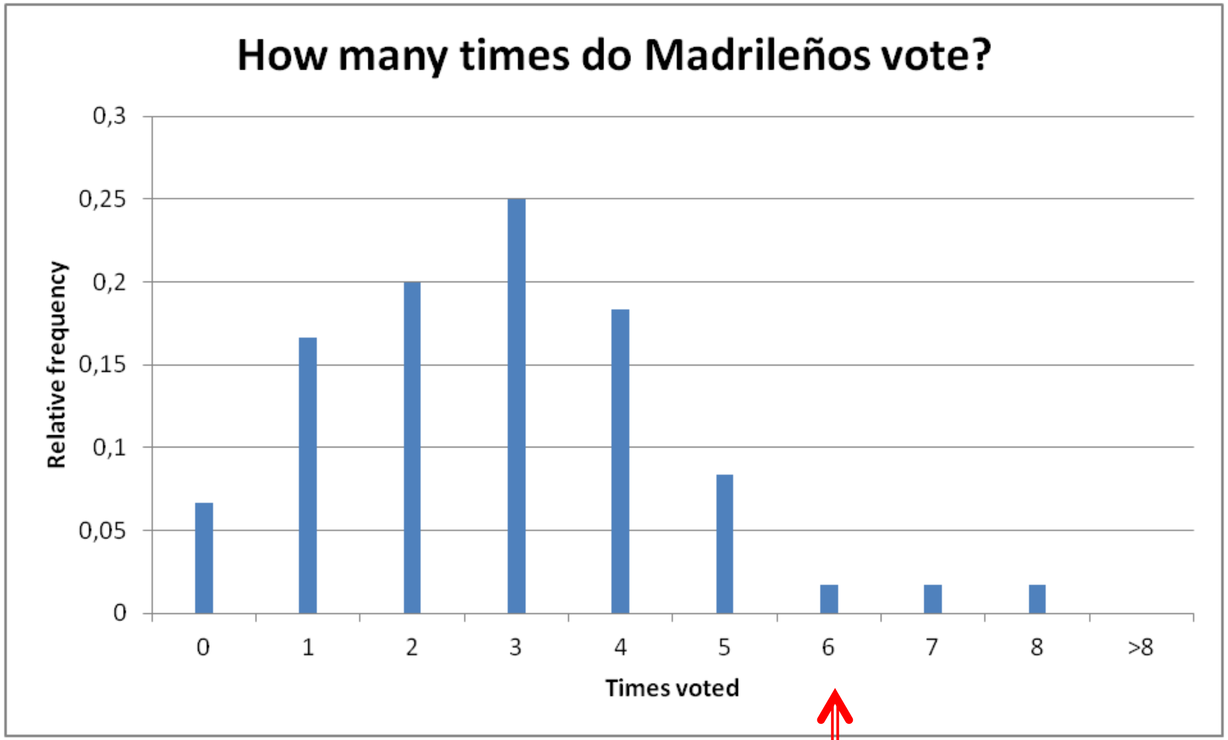
The complete table

How many people have voted less than three times?

We include an empty bar at the end →

Times voted	Absolute frequency	Cumulative frequency	Relative frequency	Cumulative relative frequency
0	4	4	0,0667	0,0667
1	10	4+10 = 14	0,1667	14/60 = 0,2333
2	12	4+10+12 = 26	0,2000	0,4333
3	15	41	0,2500	0,6833
4	11	52	0,1833	0,8667
5	5	57	0,0833	0,9500
6	1	58	0,0167	0,9667
7	1	59	0,0167	0,9833
8	1	60	0,0167	1,0000
>8	0	60	0,0000	1,0000
Total	60		1,0000	

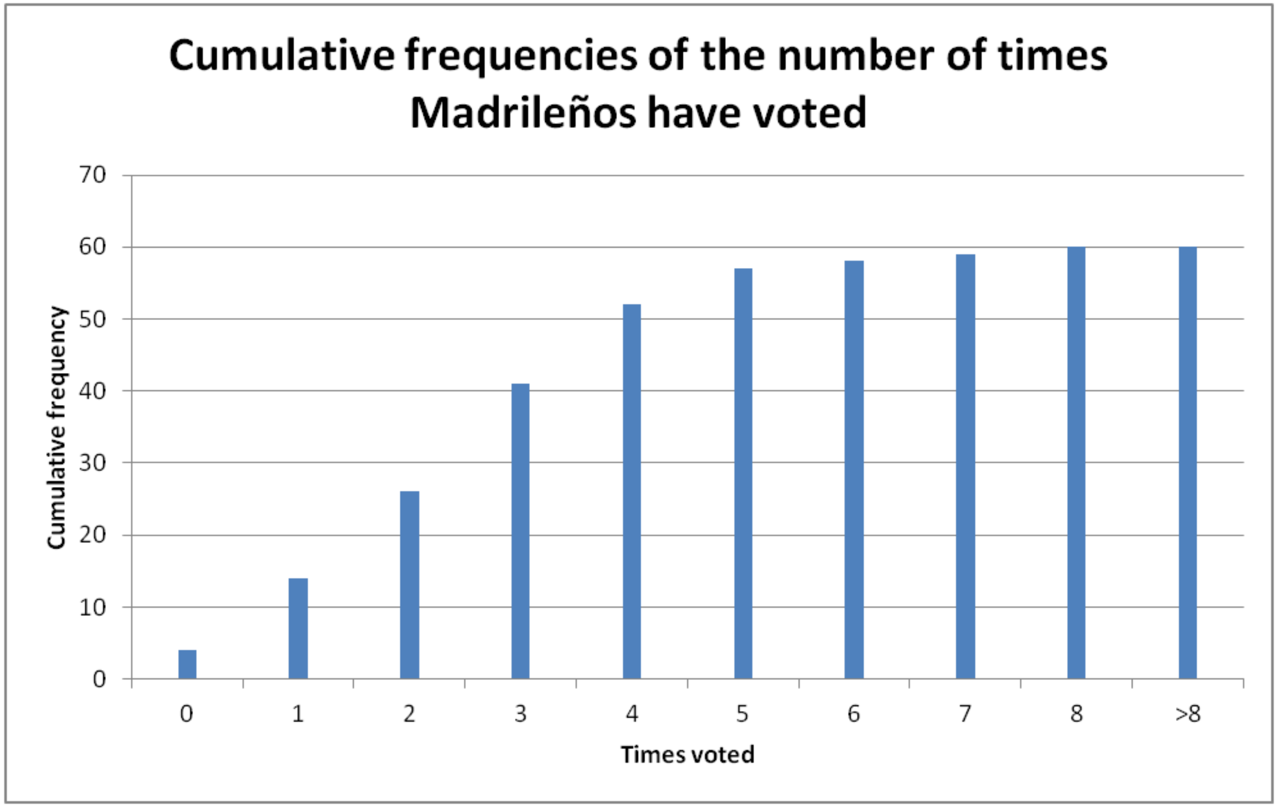
The bar chart



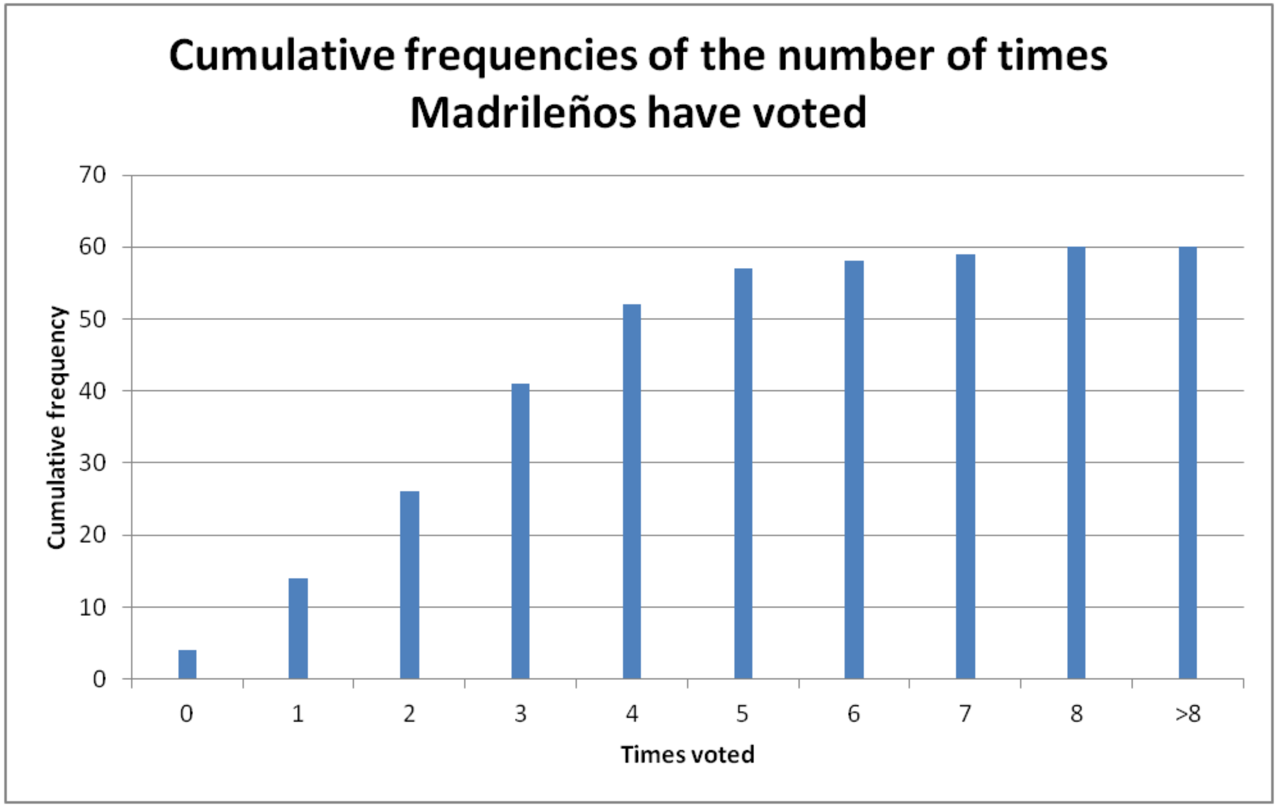
Thin bars!

What does the shape of the graph tell us?

The cumulative frequency bar chart



The cumulative frequency bar chart



Continuous data: the histogram

- When data are discrete (with few different values) it is straightforward to calculate a frequency table.
- With continuous data, it does not make sense to have a separate category for each data value.

Why?

Money received by 36 Madrid municipalities in 1995 (1000s of PTAS)

114579 73896 59003 86165 53428 93844 61536 90628 49501

56767 78063 87750 82409 107664 60479 88872 66325 78268

38360 82436 83531 81364 63210 112842 56206 59052 52660

45000 91562 66308 50397 79964 65369 71803 60108 49264

http://wwwmadridorg/iestadis/fijas/estructu/general/territorio/im00_23htm

How many bars and where to start?

How many bars?

Group the data into approximately $\sqrt{N}$ bars.

($N = 36$, $\sqrt{N} = 6$)

How should we choose the bar widths?

Try to use round numbers for bar widths, start and end points.

(min = 38360, max = 114579)

(start = 30000, end = 120000, width = 15000)

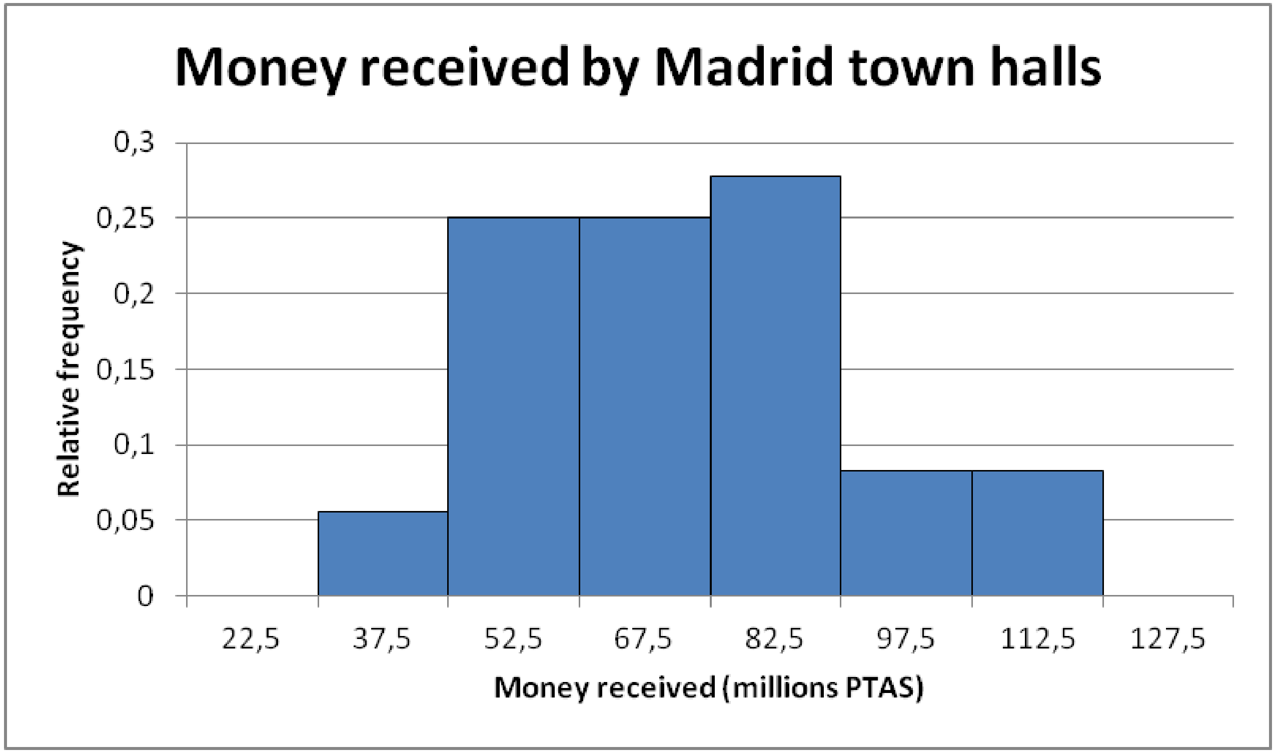
Could we use other values?

The frequency table

Take care with the end points!

Money received (millions of PTAS)	Interval centre	Abs. freq.	Cum. abs. freq.	Rel. freq.	Cum. rel. freq.
≤ 30	22,5	0	0	0	0
(30,45]	37,5	2	2	0,056	0,056
(45,60]	52,5	9	11	0,25	0,306
(60,75]	67,5	9	20	0,25	0,556
(75,90]	82,5	10	30	0,278	0,833
(90,105]	97,5	3	33	0,083	0,917
(105,120]	112,5	3	36	0,083	1
> 120	127,5	0	36	0	1
Total		36		1	

The histogram



Thick bars!

What can we say about the shape of the data?

What happens if we change the number of bars?

Variable bar widths

The table shows weekly cannabis consumption for a sample of US users.

g/week Interval [ )		Centre	Abs. freq.	Rel. freq.
0	3	1,5	94	0,178
3	11	7	269	0,509
11	18	14,5	70	0,132
18	25	21,5	48	0,091
25	32	28,5	31	0,059
32	39	35,5	10	0,019
39	46	42,5	5	0,009
46	74	60	2	0,004
74	+	90	0	0
	Total		529	1

What is wrong with graphing this directly?

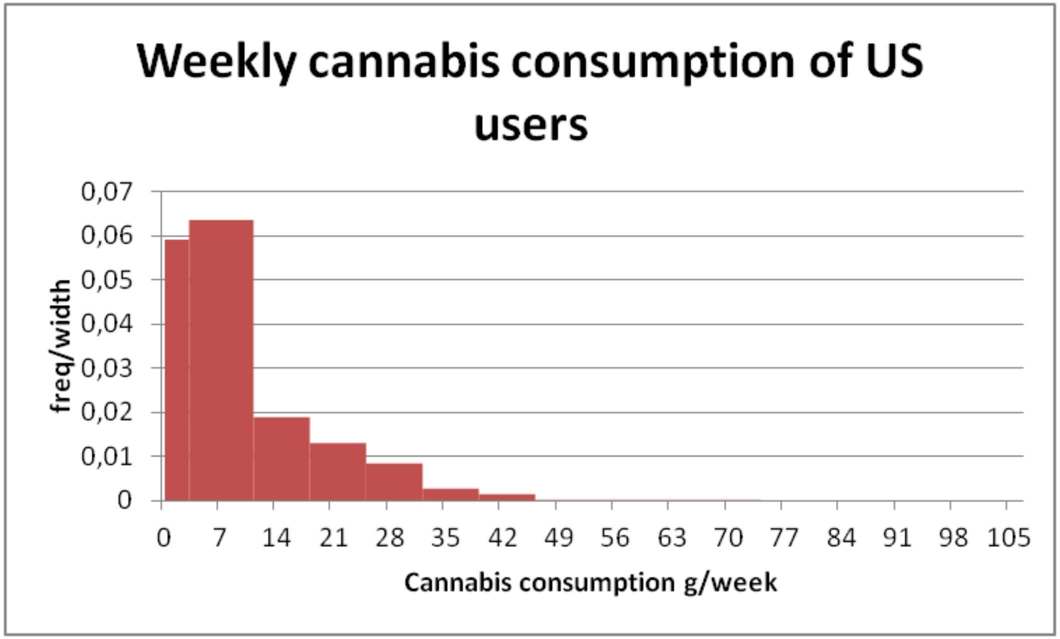
Adjusting the height

g/week Interval [ )		Centre	Abs. freq.	Rel. freq.	Height
0	3	1,5	94	0,177693762	0,059
3	11	7	269	0,508506616	0,064
11	18	14,5	70	0,132325142	0,019
18	25	21,5	48	0,09073724	0,013
25	32	28,5	31	0,058601134	0,008
32	39	35,5	10	0,018903592	0,003
39	46	42,5	5	0,009451796	0,001
46	74	60	2	0,003780718	1E-04
74	+	90	0	0	0
	Total		529	1	

We use the formula:

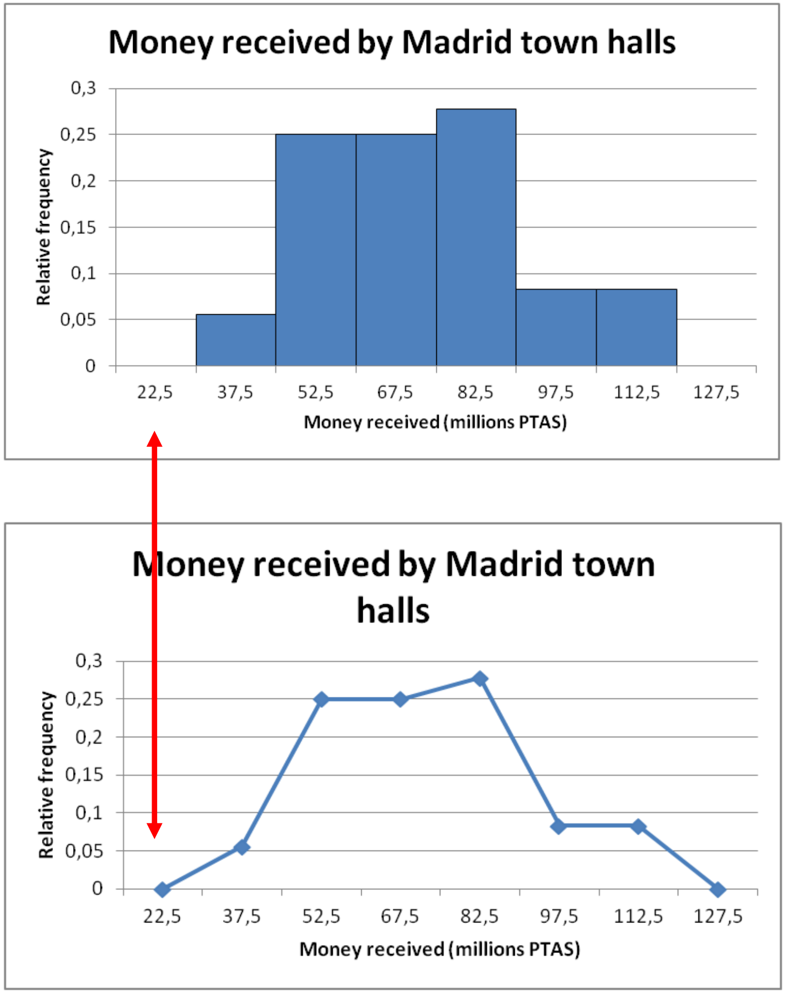
height = frequency / width.

The histogram



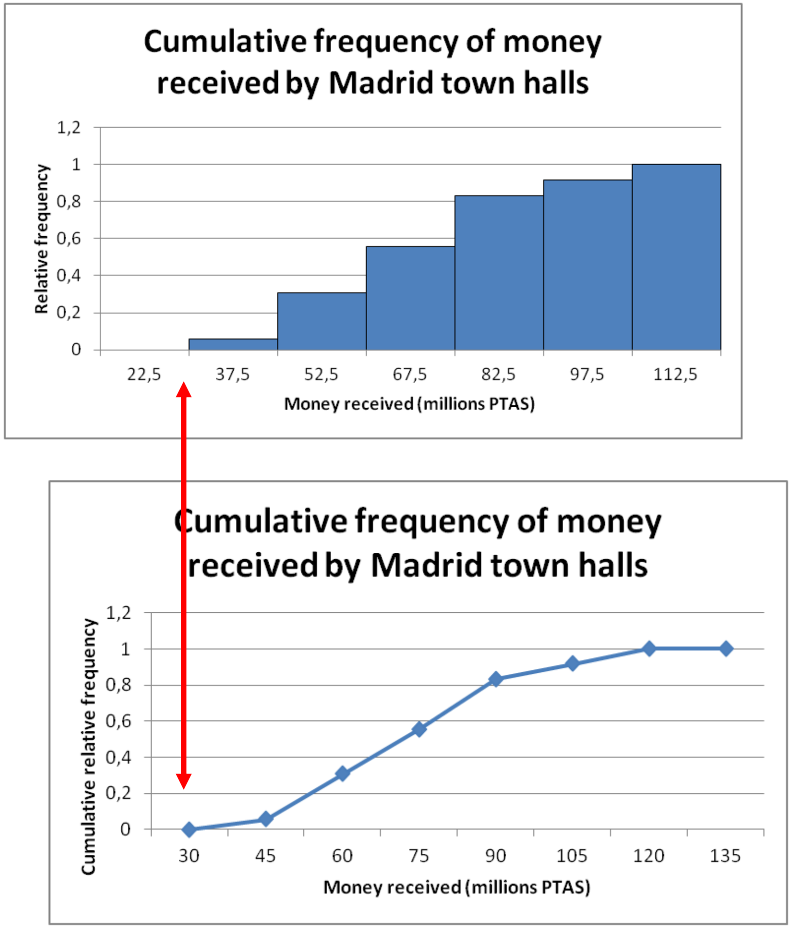
The data are very skewed to the right.

The frequency polygon



This is a smoothed histogram. Each bar is joined at the centre.

The frequency polygon with cumulative frequencies



Join up at the ends of the bar intervals.

Exercise

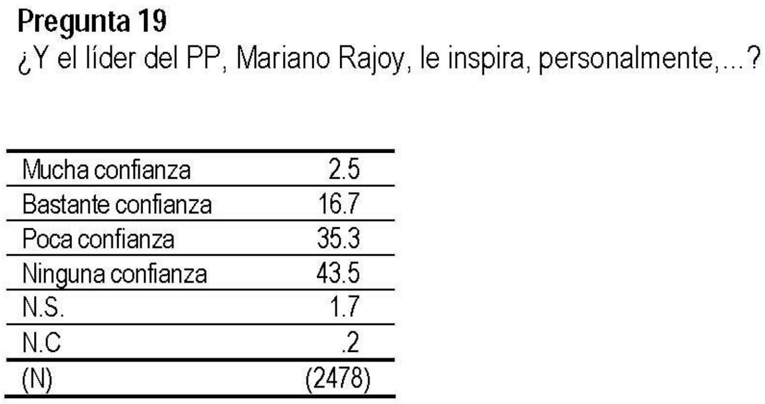
The 40 students in a statistics class rate their lecturer from 1 (extremely boring) to 5 (fantastic). The table partially shows the survey results.

Evaluation	Absolute frequency	Relative frequency
1		0,05
2		
3	5	
4	9	
5	19	
TOTAL		

Complete the table.

Exercise

The following table comes from the CIS survey of January 2011. The values are given as (approximate) percentages of a total number of 2478 respondents.



Pregunta 19

¿Y el líder del PP, Mariano Rajoy, le inspira, personalmente,...?

Mucha confianza	2.5
Bastante confianza	16.7
Poca confianza	35.3
Ninguna confianza	43.5
N.S.	1.7
N.C	.2
(N)	(2478)

Which of the following affirmations is correct?

a) The number of respondents who have a lot of confidence (*mucha confianza*) in the Mariano Rajoy is approximately 619.

b) Approximately 1953 of the respondents have little or no confidence (*poca o ninguna confianza*) in the leader of the PP.

c) The relative frequency of respondents who don’t know (*NS*) or don’t reply (*NC*) is 0.19.

d) None of the above.

Exercise

The following pie chart shows the distribution of the autonomous communities visited by foreign tourists.

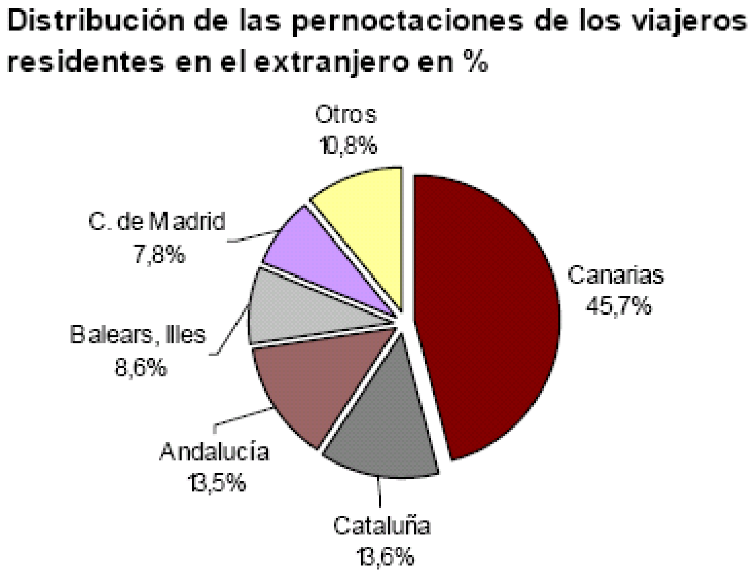
Which of the following is the correct response?

a) The percentage of tourists who visit the islands is lower than the percentage for the rest of the destinations.

b) The percentage of tourists who visit the islands is higher than the percentage for the rest of the destinations.

c) Cataluña and the Comunidad de Madrid are the communities with the highest percentages of foreign tourists.

d) None of the above.



Exercise

The following pie chart concerns the voting concerns of students at the University of Houston before the 2010 elections.

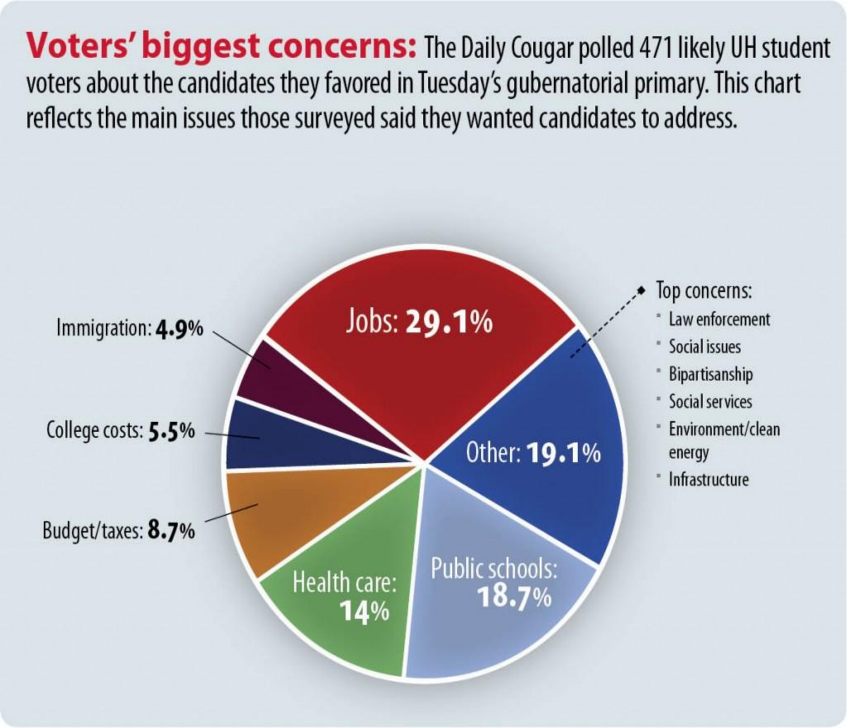
Which of the following affirmations is correct?

a) 160 students said that the main issues were *Jobs* or *Immigration*.

b) 327 students said that the main issues were *Public schools* or *Health care*.

c) 25 students said that the main issue was *Other*.

d) 259 students said that the main issue was *College costs*.



Voters' biggest concerns: The Daily Cougar polled 471 likely UH student voters about the candidates they favored in Tuesday's gubernatorial primary. This chart reflects the main issues those surveyed said they wanted candidates to address.

THE DAILY COUGAR