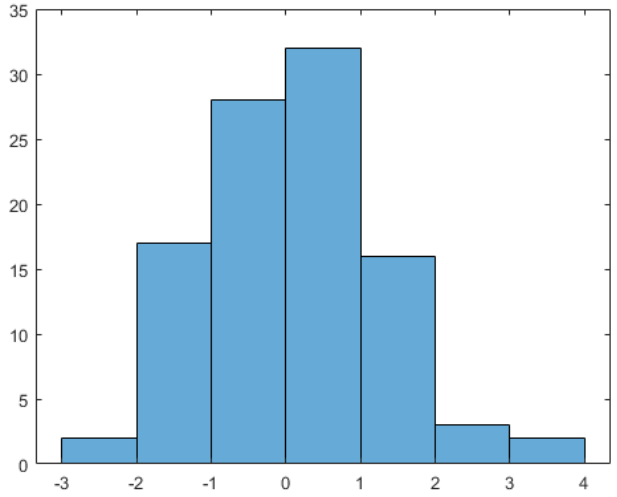
Class 3: Analysis of univariate data: graphics for quantitative data



Recommended reading:

Have a look at the Wikipedia page on histograms.

Class 3: Analysis of univariate data: graphics for quantitative data

Just like for qualitative data, the first step with quantitative data is to set up a frequency table. If the data are discrete, this is easy.

SAMPLE: 60 adult madrileños
VARIABLE: Number of times you have voted in the municipal elections
OBJECTIVE: Classification and representation of information.

3 3 3 4 1 2 4 5 2 3 1 1 3 8 4 1 3 4 2 5 0 0 5 4 2 1 2 3 3 2

1 4 3 2 3 5 0 6 3 1 3 5 4 1 4 1 2 4 4 3 3 0 7 2 2 1 3 4 2 2

The frequency table

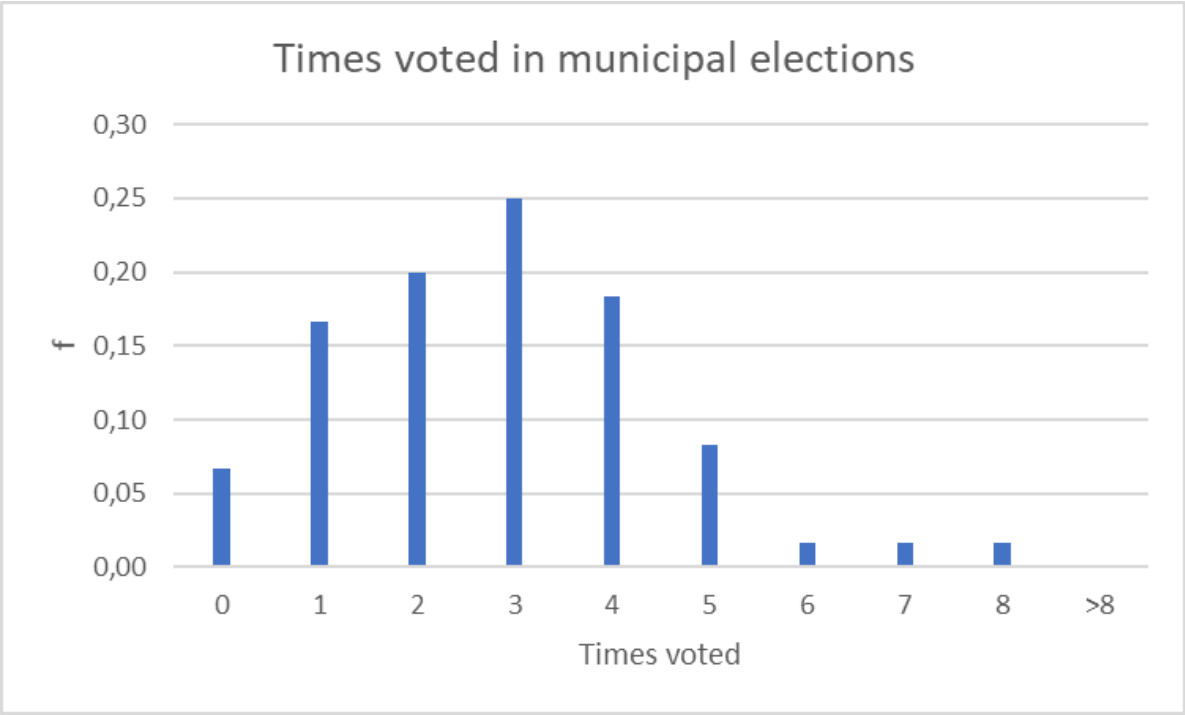
Times voted	Absolute frequency (n)
0	4
1	10
2	12
3	15
4	11
5	5
6	1
7	1
8	1
>8	0
Total	60

Include an empty bar

The extended frequency table

Times voted	Absolute frequency (n)	Cumulative frequency (N)	Relative frequency (f)	Cumulative relative frequency (F)
0	4	4	4/60 = 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	——

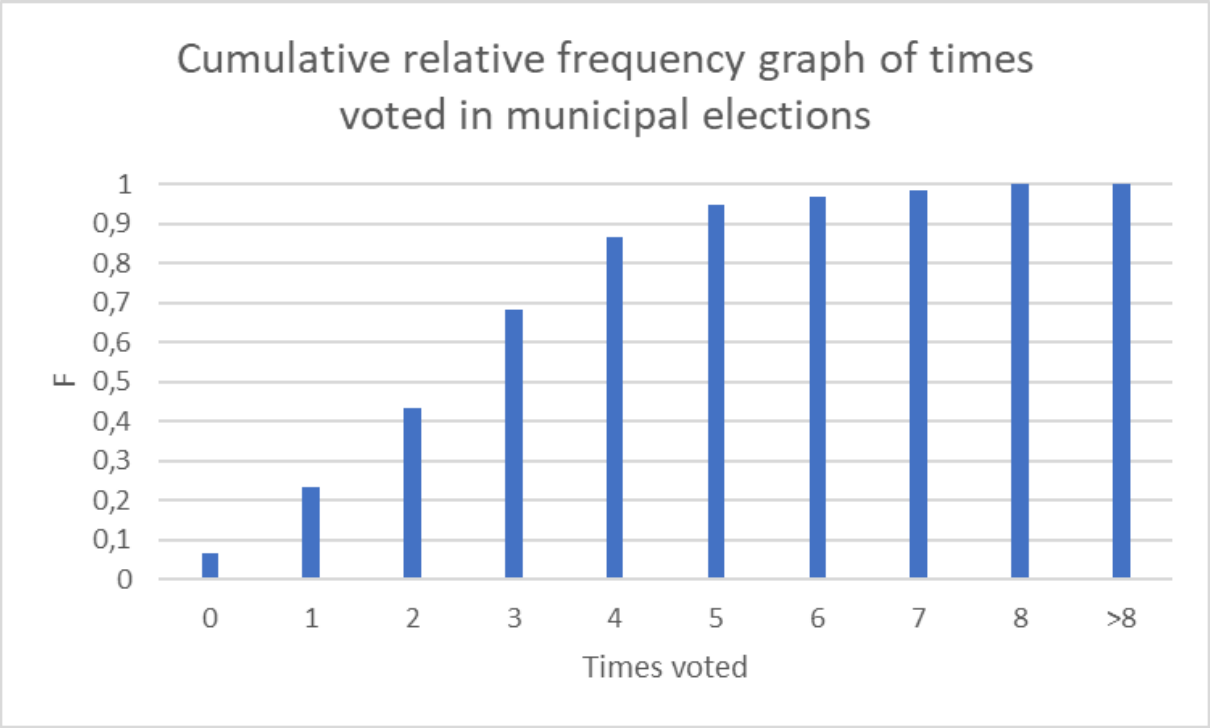
The bar chart



What are the typical values of the data?

What other features can you observe?

Cumulative frequency bar chart



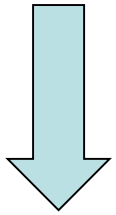
Constructing a frequency table with continuous data

Yearly spending of different town halls (€)

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

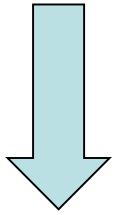
N=36
Minimum = 38360
Maximum = 114579

A bar diagram using each individual value doesn't make sense



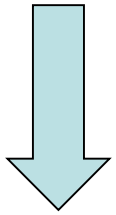
We need to group the data

How many intervals should we use?



Approximately $\sqrt{N}$

Where should we start and how wide should the intervals be?



Try to use round numbers

The frequency table

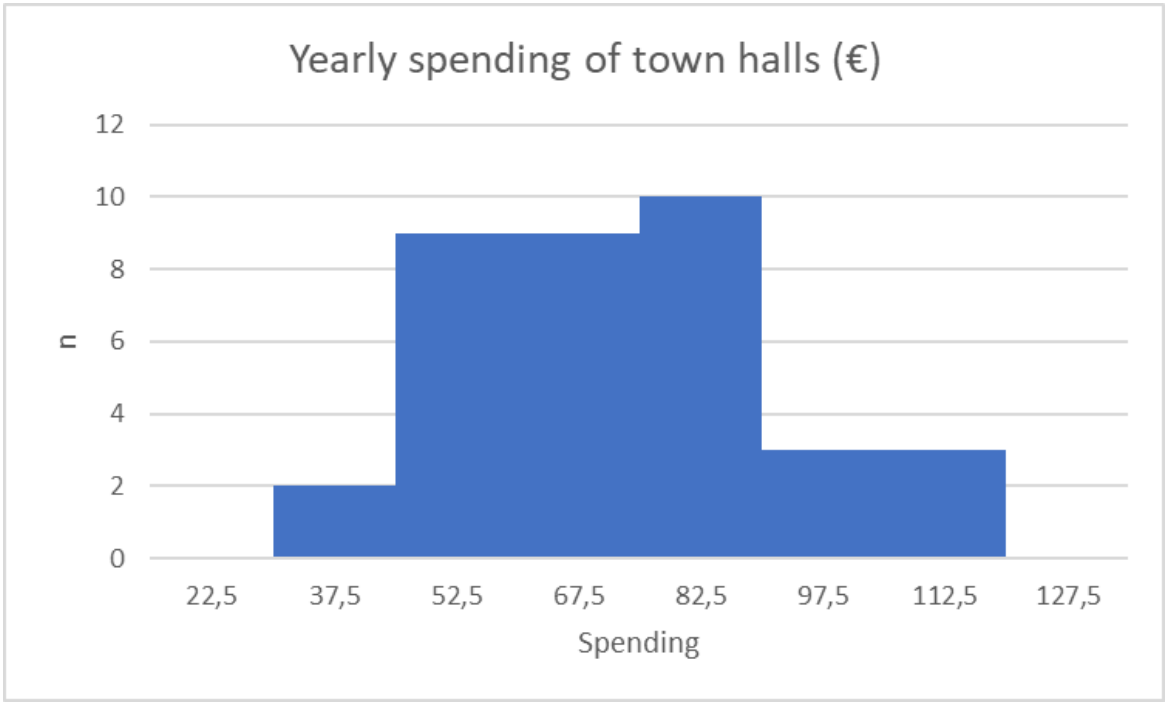
Include an empty bar

Data ≤ 45

Data > 45

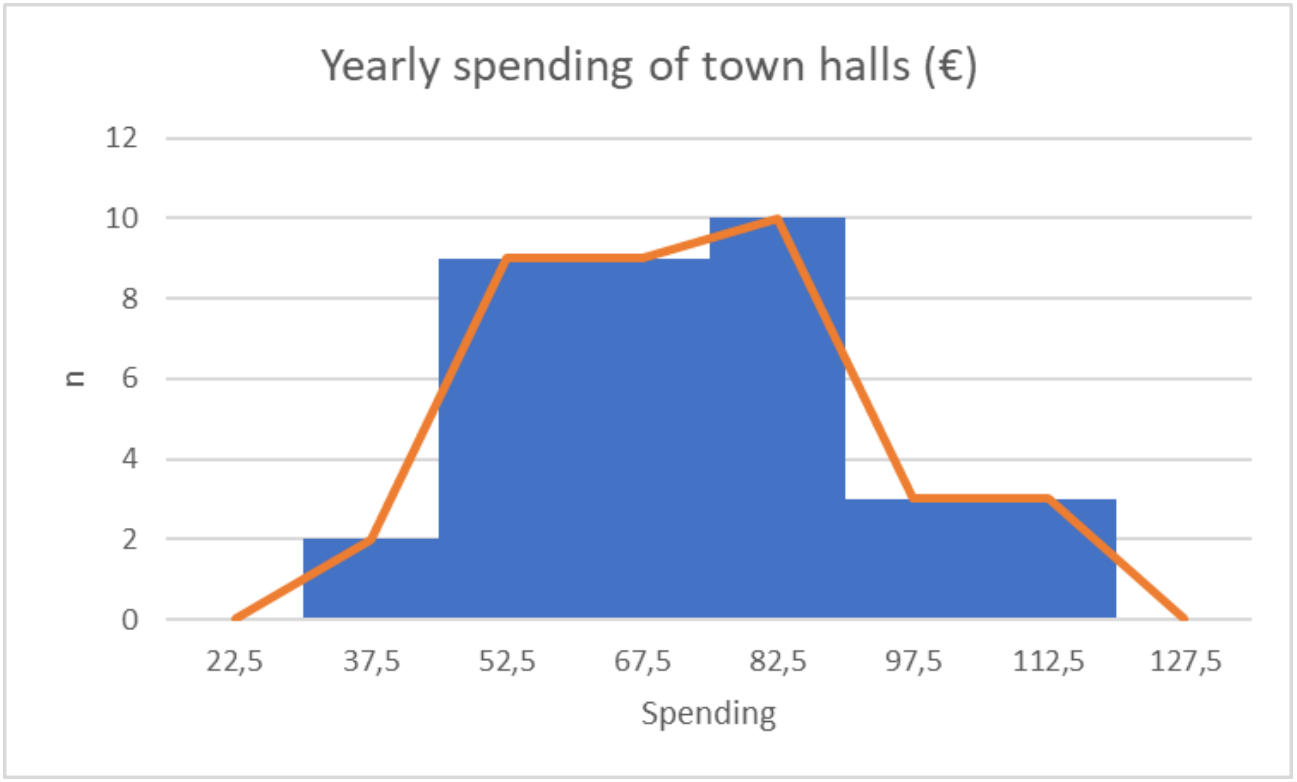
Yearly spending (€)	Absolute frequency
≤ 30	0
(30,45]	2
(45,60]	9
(60,75]	9
(75,90]	10
(90,105]	3
(105,120]	3
> 120	0
Total	36

The histogram



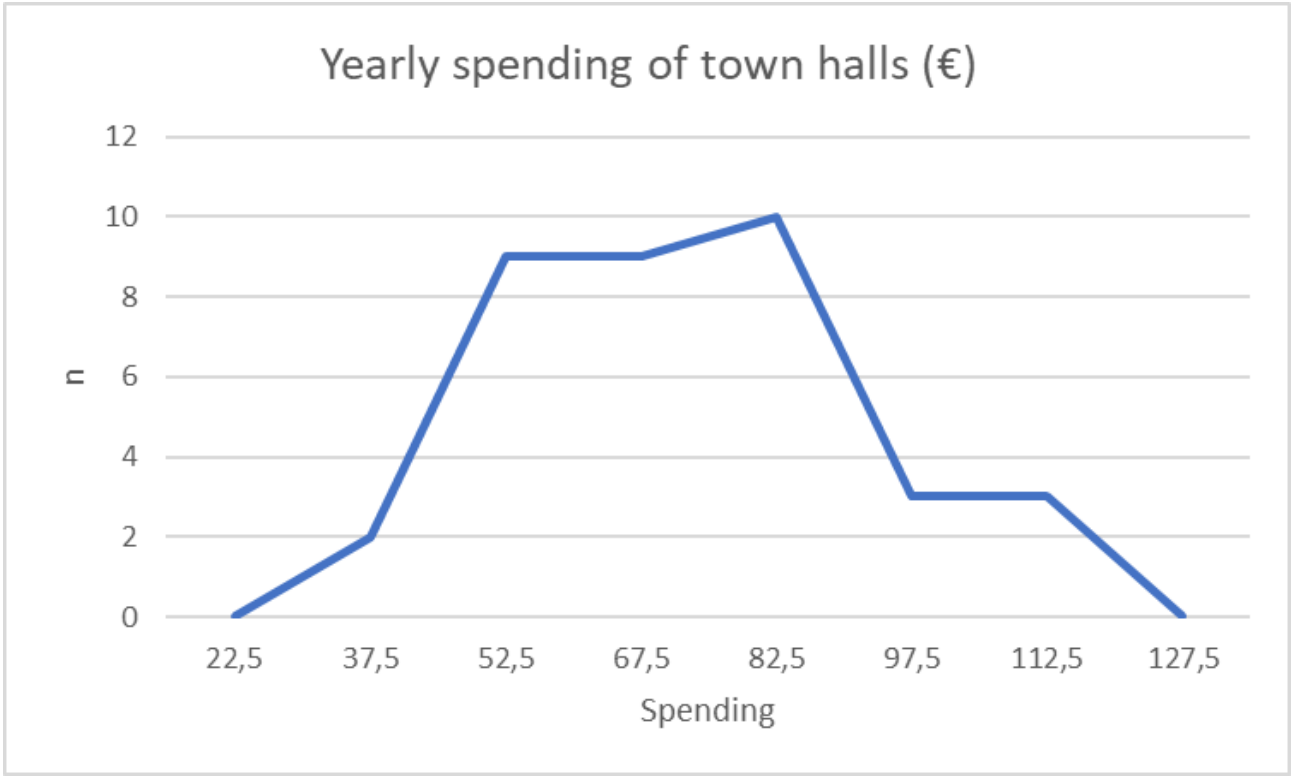
What can we say about the form of the data?

The frequency polygon



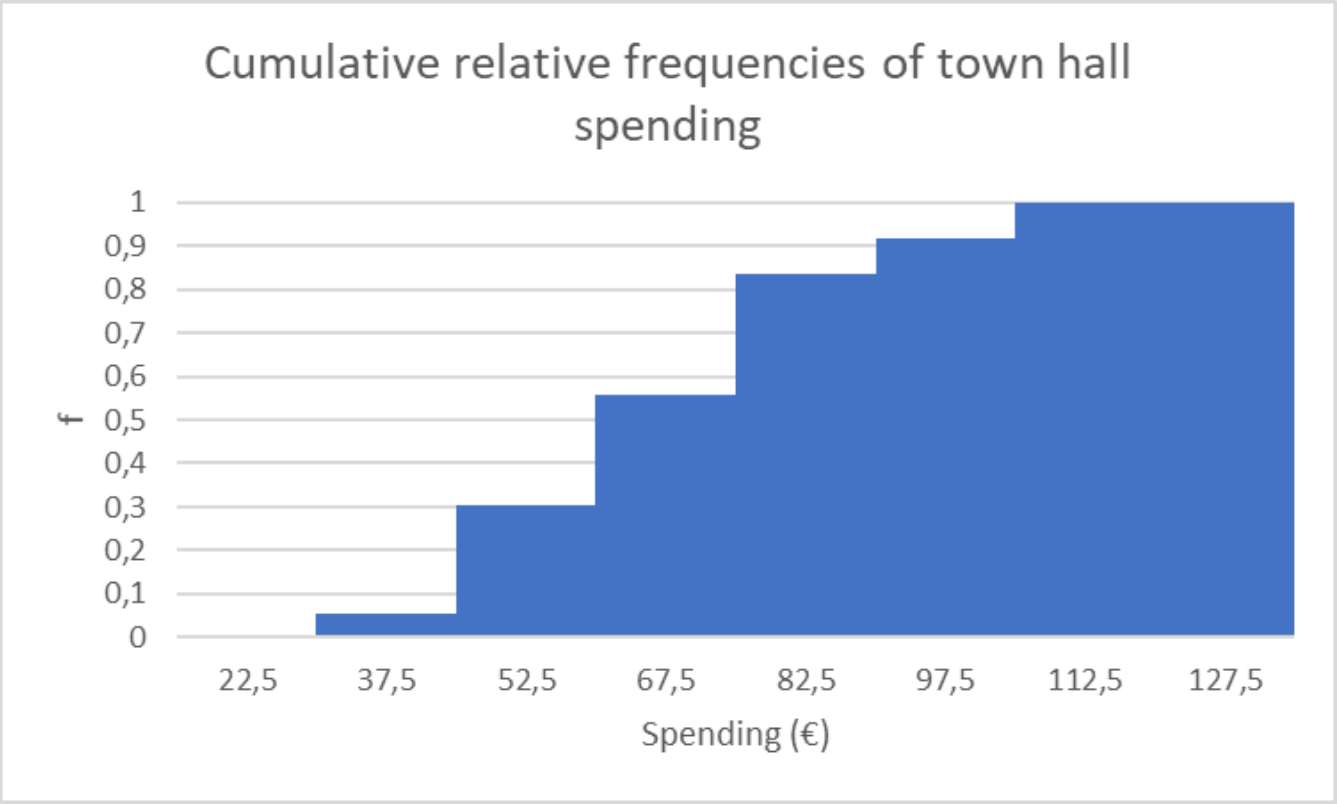
Lines joined at centre of each Interval.

The frequency polygon

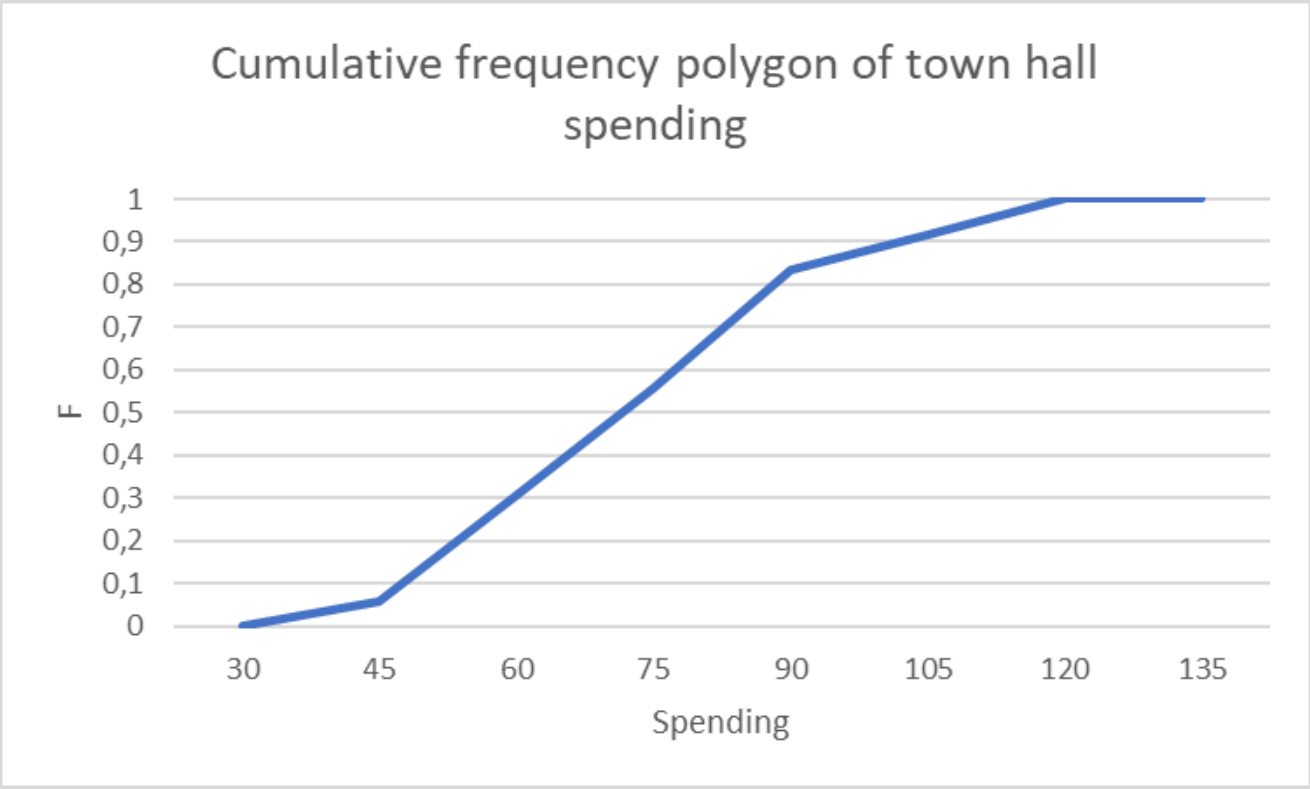


The frequency polygon is a “smoothed” histogram.

Histogram and frequency polygon for cumulative data



Histogram and frequency polygon for cumulative data



Lines joined from start to end of each Interval.

What happens if we change the number of bars in a histogram?

Try playing with this example

Small changes frpm the optimum selection make little difference to the overall shape.

Big changes can affect the shape dramatically.

What happens if we group the data in intervals of different widths

Sometimes, we have very skewed data where many values are concentrated in a small área and a few values are much further away.

Then it is often convenient to group the data into bars of different widths.

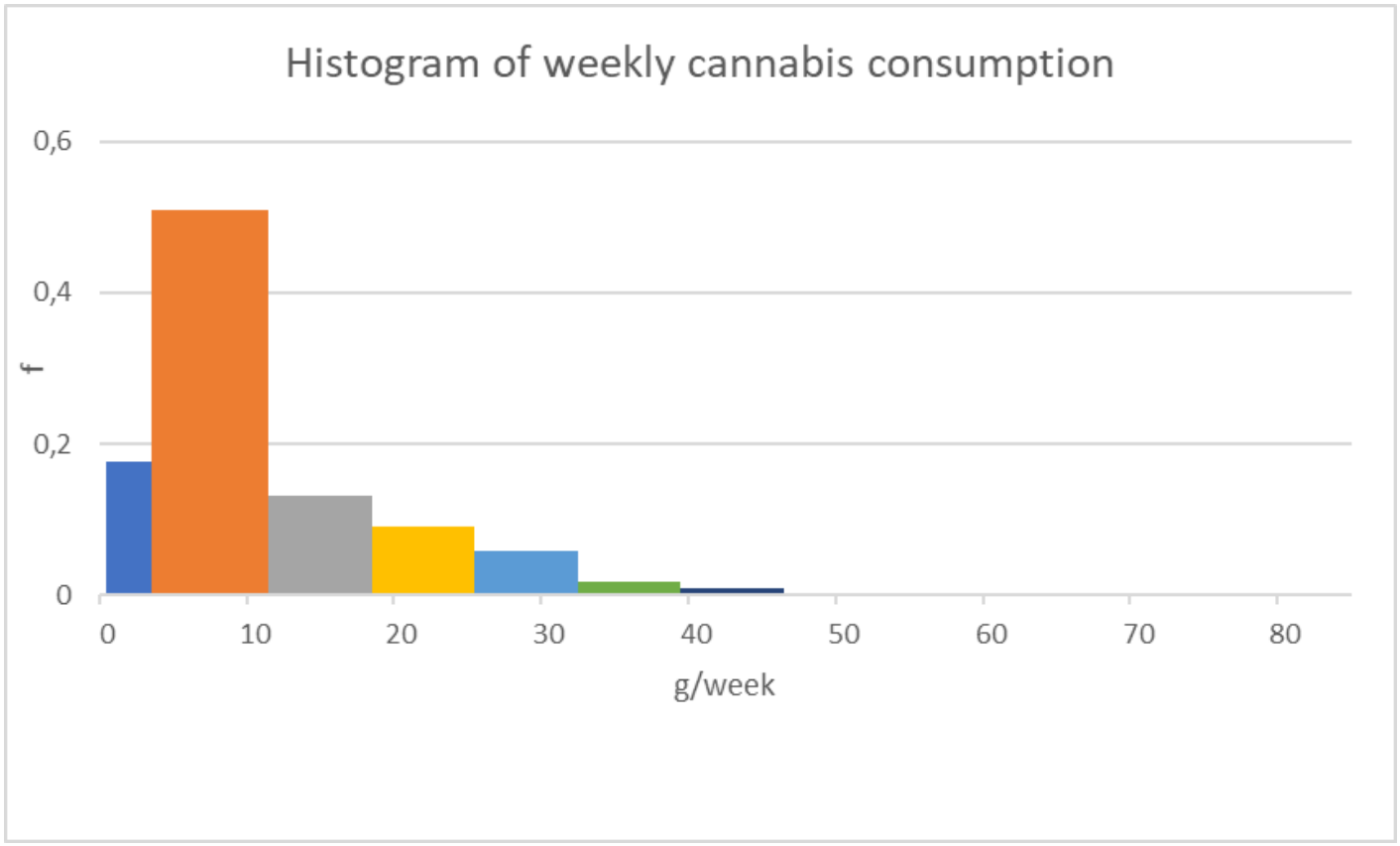
How should we draw the histogram in this case?

Example: cannabis consumption

The following data comes from a study of cannabis users and reflects their weekly consumption in grams..

Interval of consumption in g/week			
Lower limit	Centre	Upper Limit	Absolute frequency (n)
0	1,5	3	94
3	7	11	269
11	14,5	18	70
18	21,5	25	48
25	28,5	32	31
32	35,5	39	10
39	42,5	46	5
46	60	74	2
74		∞	0

Histogram with relative frequencies as heights



What impression does this give?

Do you think it looks right?

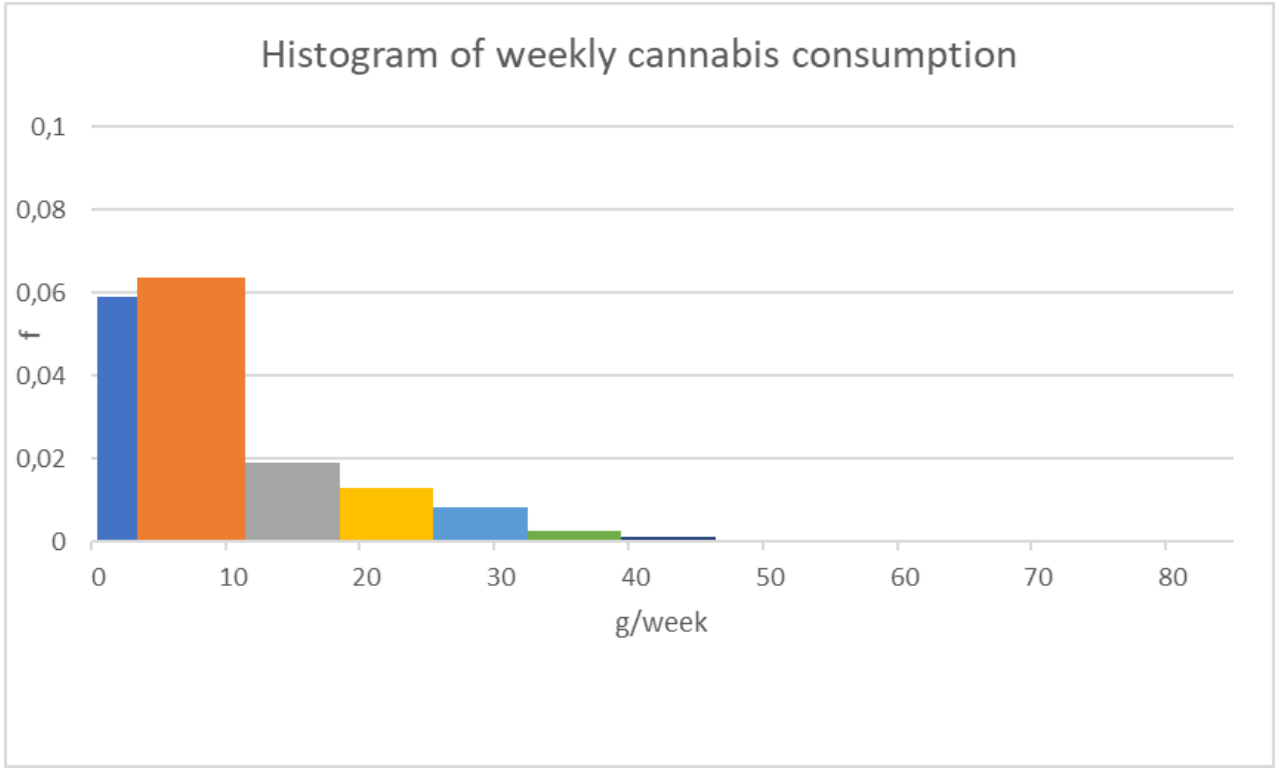
Calculating the correct heights

We visualize area not height.

If area = frequency, then height = frequency/width.

Interval of consumption in g/week						
Lower limit	Centre	Upper Limit	Width	Absolute frequency (n)	Relative frequency (f)	Height
0	1,5	3	3	94	0,178	0,059
3	7	11	8	269	0,509	0,064
11	14,5	18	7	70	0,132	0,019
18	21,5	25	7	48	0,091	0,013
25	28,5	32	7	31	0,059	0,008
32	35,5	39	7	10	0,019	0,003
39	42,5	46	7	5	0,009	0,001
46	60	74	28	2	0,004	0,000
74		∞		0	0,000	0,000
			Total	529	1,000	

The correct histogram



Observe how the form of the graph has changed.

What is the area of the graph?

Would it change the form of the graph if we did this for a standard histogram?

Is there a relation with cumulative frequency?

Exercise

The number of judicial cases open against various politicians is illustrated in the following table:

Open cases	Frequency
x_i	n_i
0	5
1	18
2	41
3	28
4	8

Construct an appropriate graphic to represent these data.

Exercise

The following table shows the number of boyfriends / girlfriends in the last year as reported in the class questionnaire:

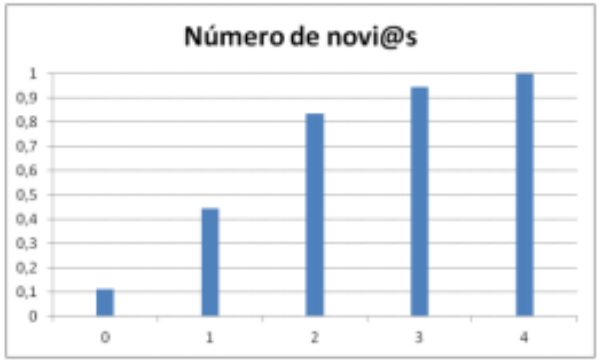
Boy/girlfriends	Frequency
0	2
1	6
2	7
3	2
4	1

What is the proportion of the class that have had two or more boy / girlfriends?

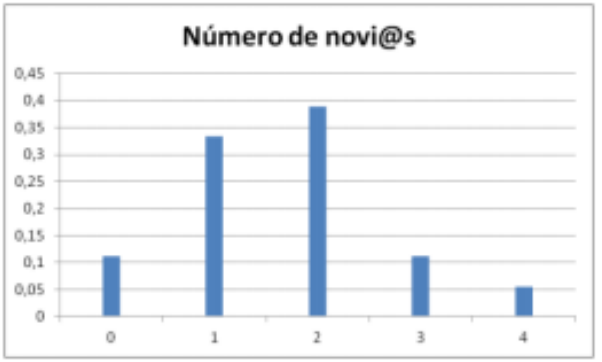
Exercise

Which of the following cumulative frequency graphs is the correct one to represent these data?:

a)



b)



c)

d)

Exercise

Impact factor	Frequency
(0, 0.4]	21
(0.4, 0.8]	30
(0.8, 1.2]	30
(1.2, 1.6]	26
(1.6, 2]	18
(2, 2.5]	16
(2.5, 3]	14
(3, 3.5]	8
(3.5, 4]	4
(4, 5]	1
(5,6]	1
>6	0

The table shows the impact factors of journals in the POLITICAL SCIENCE category of the Journal Citation Reports in 2017.

Construct an approporiate histogram for these data and briefly comment the results.