Looking at Data.

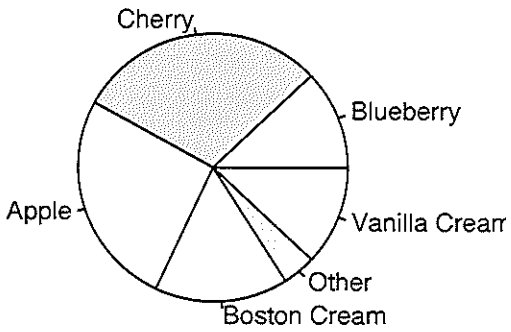
Graphical representations that can aid our understanding of the distribution of a data set.

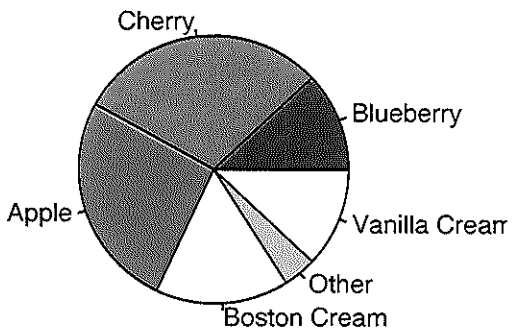
- range of values
- most likely values
- symmetry
- uni/multi-modality
- tail behaviour

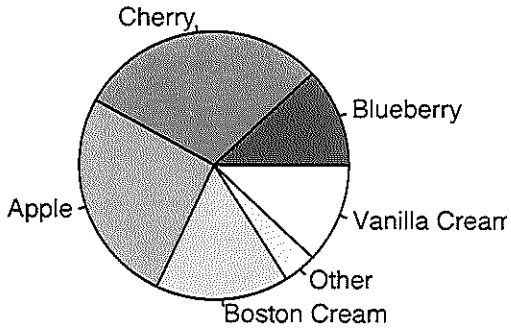
Pie charts - typically a bad way of showing data.

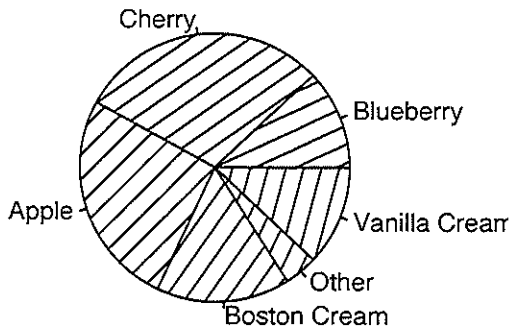
- we are good at judging linear measures not angles or areas.

Better ways : Bar chart.
Dot chart.

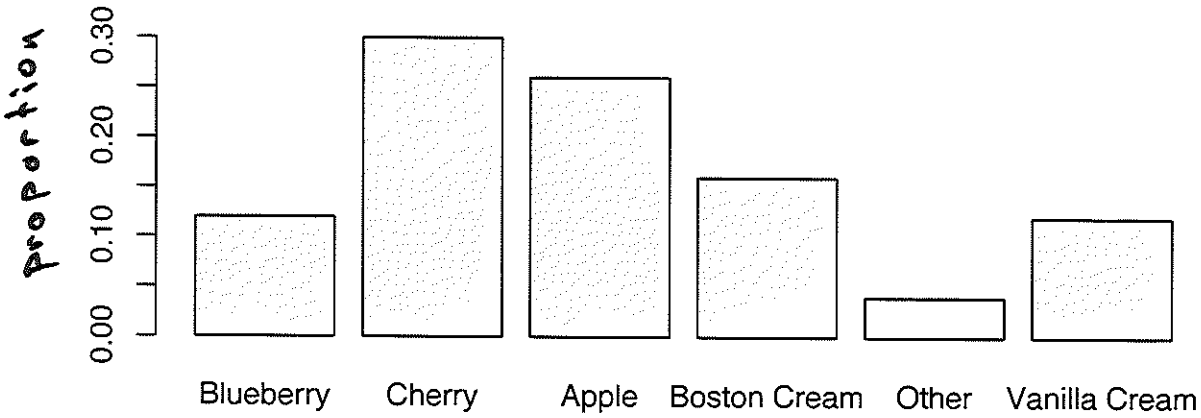

Cherry
Blueberry
Apple
Vanilla Cream
Other
Boston Cream


Cherry
Blueberry
Apple
Vanilla Cream
Other
Boston Cream


Cherry
Blueberry
Apple
Vanilla Cream
Other
Boston Cream

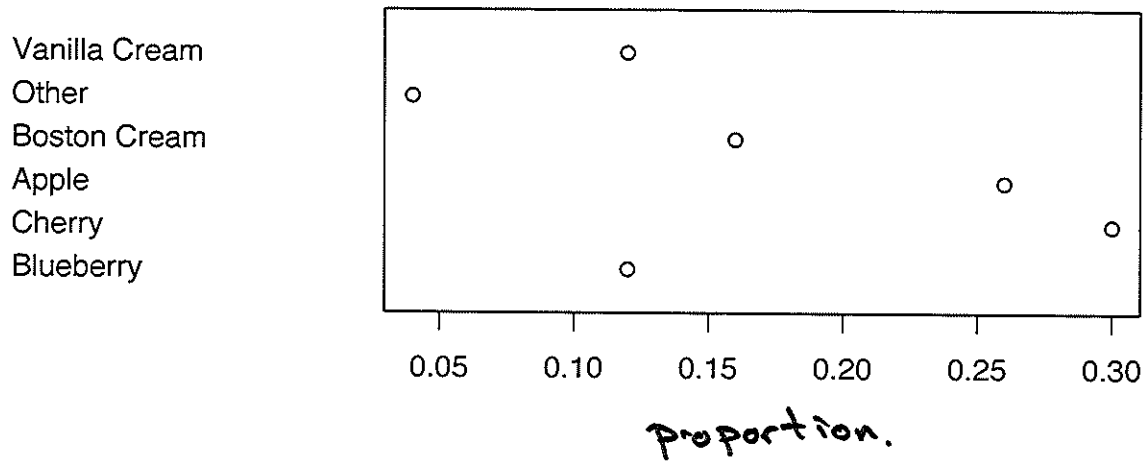

Cherry
Blueberry
Apple
Vanilla Cream
Other
Boston Cream

Bar Chart



Dot Chart

categorical variables →



Histogram - visual display of variability.

eg average length of stay in hospital.

collected from 131 hospitals.

shortest average stay 6 days

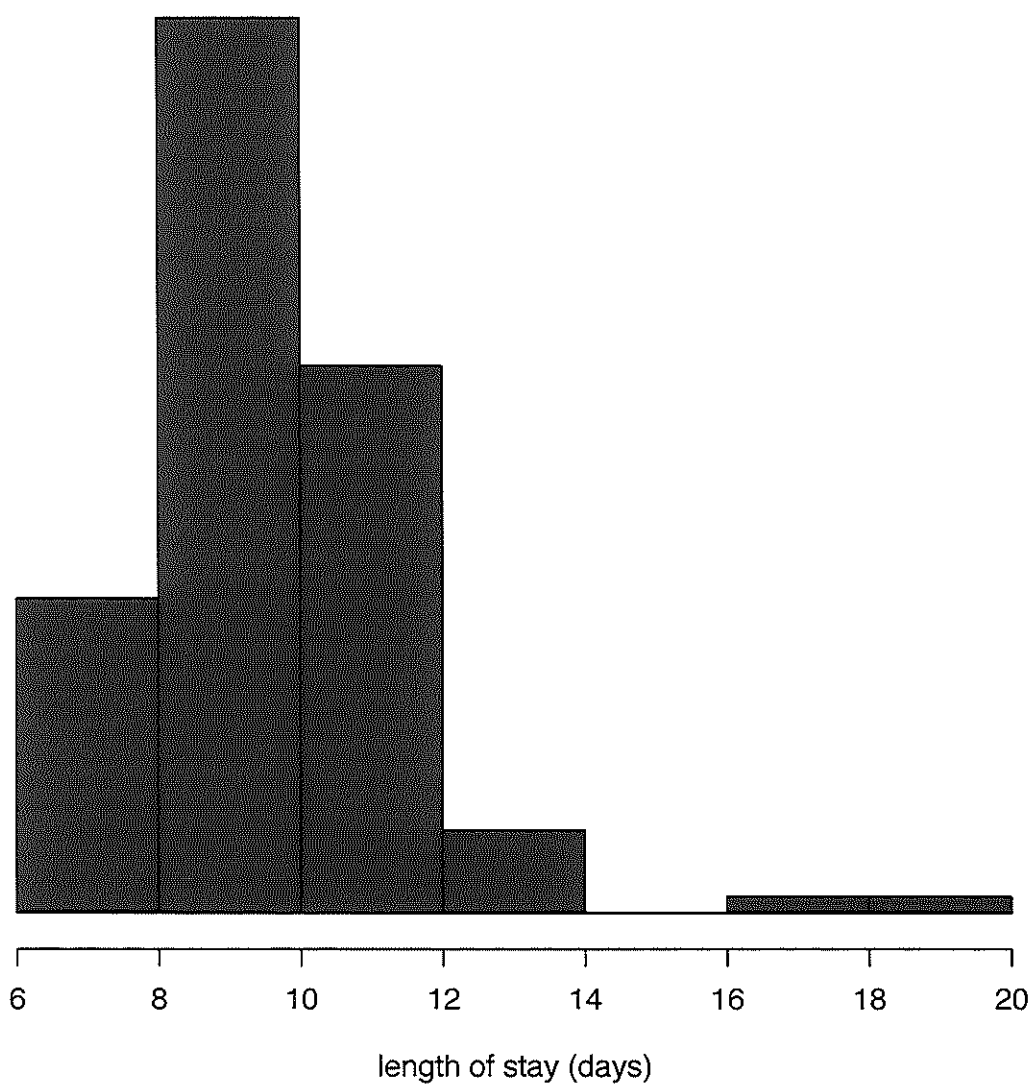
largest 20

Divide the range into class intervals.

days	# hospitals.
6 - 8	22
8 - 10	63
10 - 12	38
12 - 14	6
14 - 16	0
16 - 18	1
18 - 20	1

Draw blocks proportional to the number in each interval.

We chose equally sized class intervals. - in this case the height of the block gives all the relevant information.

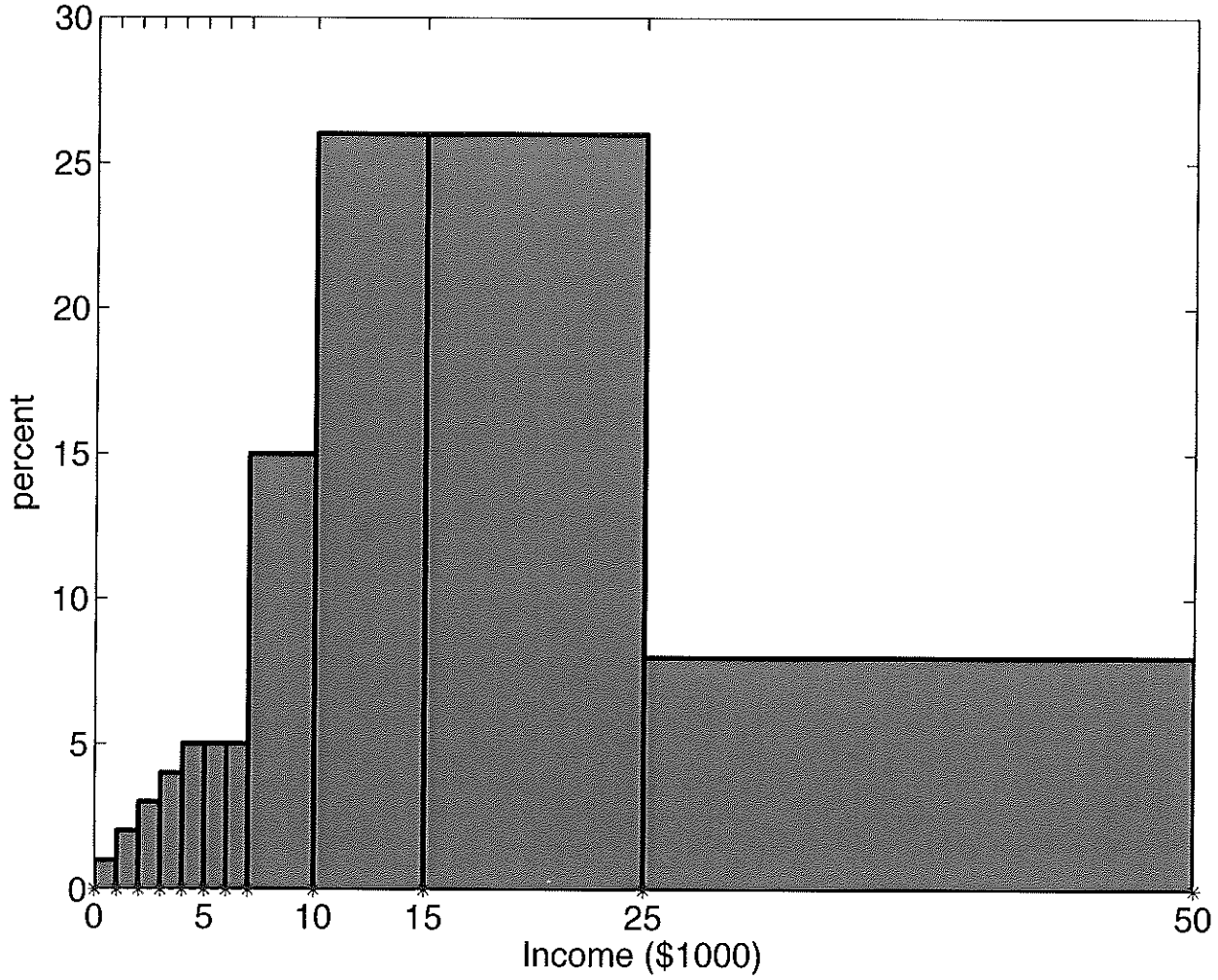

Histogram of the average length of stay in hospital
6
8
10
12
14
16
18
20
length of stay (days)

Income level in $	percent
0 – 1,000	1
1,000 – 2,000	2
2,000 – 3,000	3
3,000 – 4,000	4
4,000 – 5,000	5
5,000 – 6,000	5
6,000 – 7,000	5
7,000 – 10,000	15
10,000 – 15,000	26
15,000 – 25,000	26
25,000 – 50,000	8
50,000 and over	1

class intervals are of different widths.

↑ include the left endpoint, exclude the right endpoint

eg an income of exactly $2000 would fall into the 2000-3000 class interval.



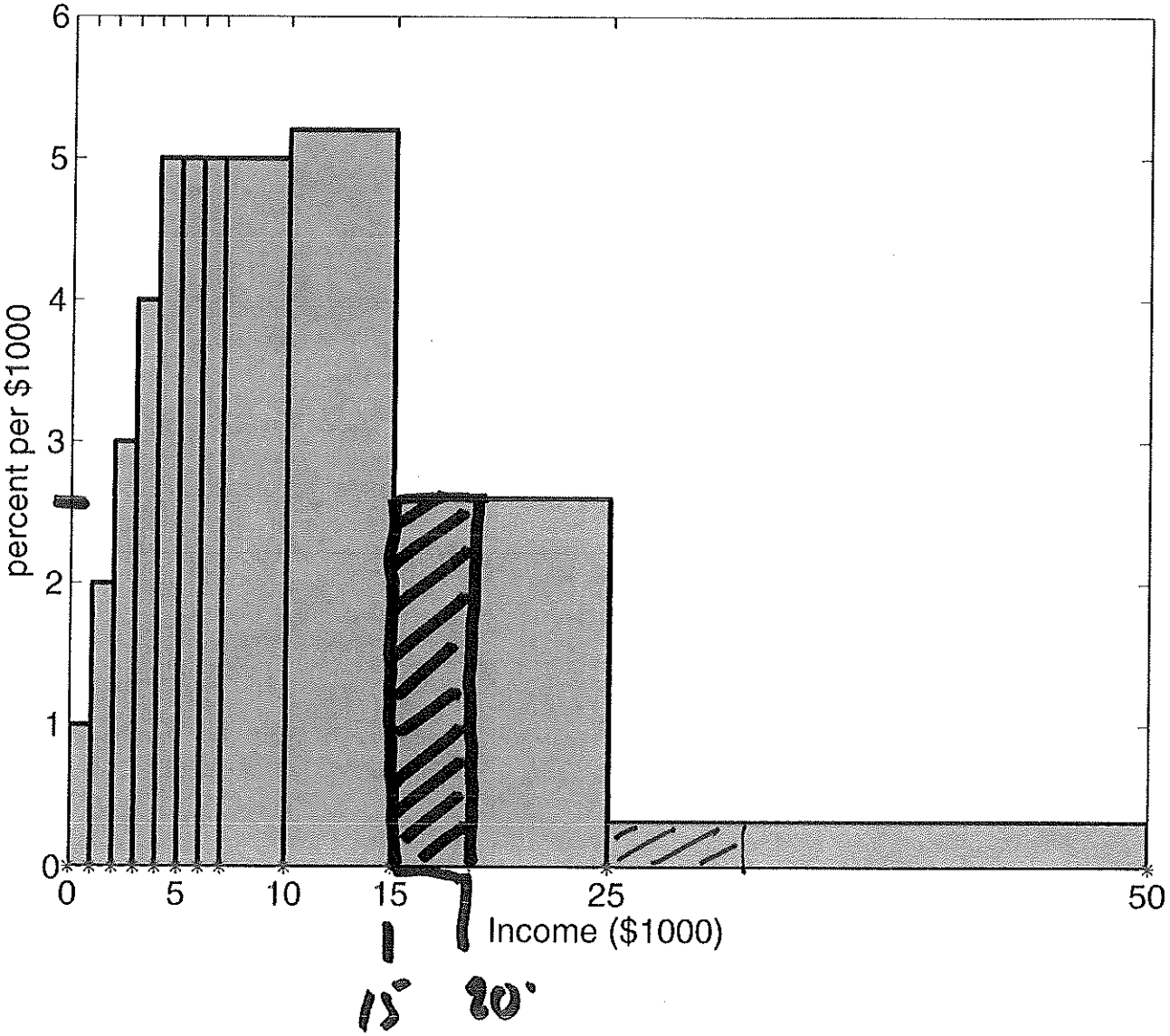
plotting bars with widths equal to the class interval + height equal to the frequency / %

can be visually very misleading.

Income level in $	percent	length (× $1,000)	height → % / interval width
0 – 1,000	1	1	1
1,000 – 2,000	2	1	2
2,000 – 3,000	3	1	3
3,000 – 4,000	4	1	4
4,000 – 5,000	5	1	5
5,000 – 6,000	5	1	5
6,000 – 7,000	5	1	5
7,000 – 10,000	15	3	5
10,000 – 15,000	26	5	5.2
15,000 – 25,000	26	10	2.6
25,000 – 50,000	8	25	.32
50,000 and over	1		

Area represents percentage.

- better visual representation
- allows us to answer questions like "What % of incomes were between $25k and $30,000"?
  - obtained by computing the blue shaded area on the histogram.



15 20

Area represents %.

What do we do when this is not the case?
When the class intervals are of varying width?

To compute the height of the bar.

take % of data in the class interval and divide by width of class interval.

What is the vertical scale?

— what are its units?

$\frac{\%}{\text{class interval width}}$ ← has units $\frac{\%}{\$1000}$.

or % per $1000.

Carpet is priced in $ per yard.

Cost for a certain number of yards is

$ per yard × # yards. → $.

$\frac{\$}{\text{yard}}$ × yards → $.

If we want to know the % between two income values, we compute the relevant area.

eg % between 15 and 20.

$= 5 \times 2.5 = 12.5\ \%$.

5: width in $1000 units

2.5: height in % / $1000

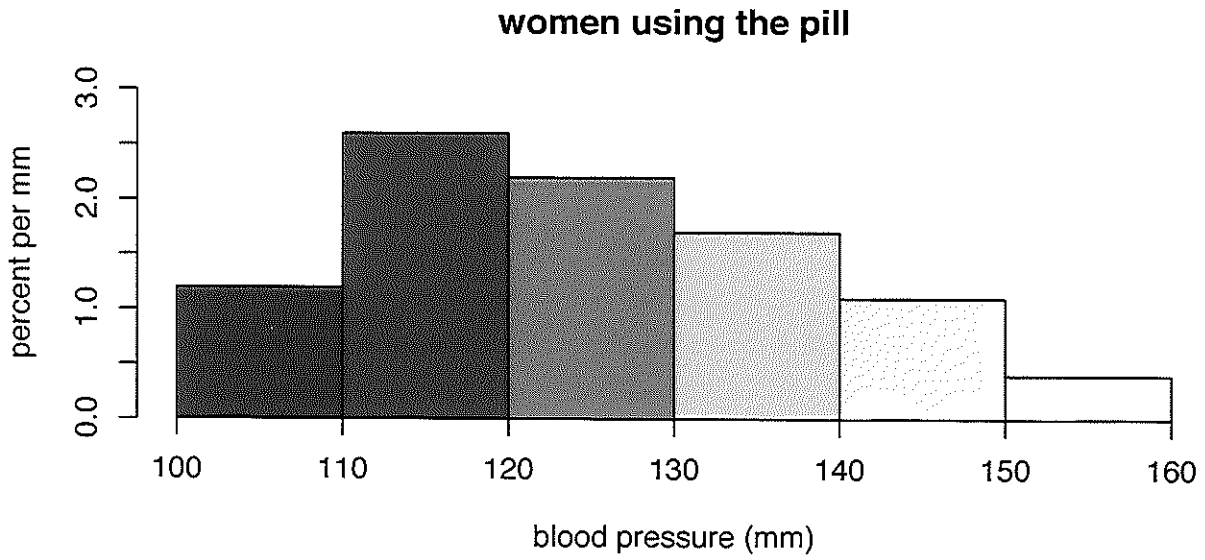
blood pressure (mm)	non users %	users %
under 100	8	6
100–110	20	12
110–120	31	26
120–130	19	22
130–140	13	17
140–150	6	11
150–160	2	4
over 160	1	2

↑

hard to tell by looking at the table whether the distributions of blood pressure measurements are different for the two groups.

From the histograms, visually we can see that the treatment group has values that are slightly higher than those for the non-treatment group.

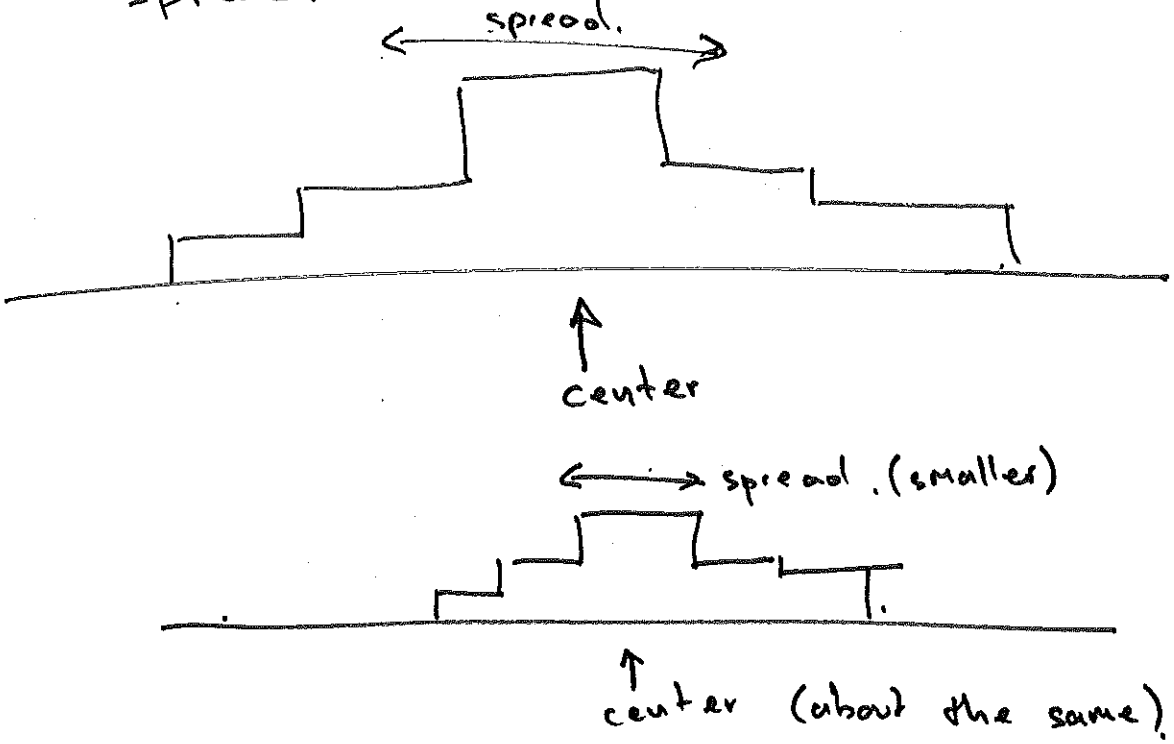

women not using the pill
percent per mm
0.0
1.0
2.0
3.0
100
110
120
130
140
150
160
blood pressure (mm)


women using the pill
percent per mm
0.0
1.0
2.0
3.0
100
110
120
130
140
150
160
blood pressure (mm)

Histogram shows the distribution of the data.

Sometimes we need to summarize the distribution by just a few numbers.

center
spread.



Measures of Center.

"average" — there are many different "averages"

→ usually this refers to the arithmetic mean (or "mean")

the average (mean) of a list of numbers equals their sum, divided by how many numbers their are.

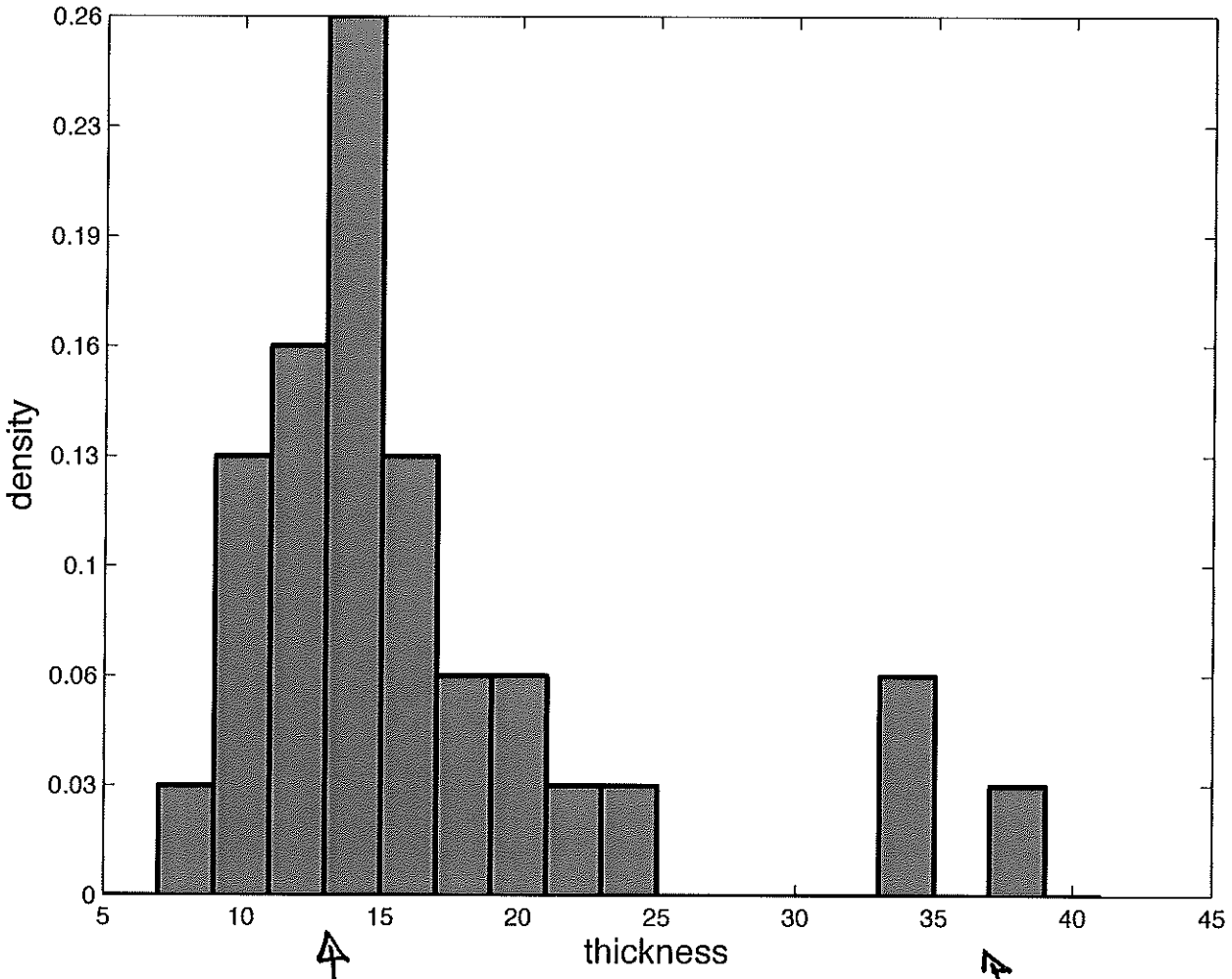
example: 36, 33, 36, 39, 37, 40

(these are the thicknesses of 6 vine leaves).

$$\text{mean} = \frac{36 + 33 + 36 + 39 + 37 + 40}{6}.$$

$$= \frac{221}{6} \approx 37.$$

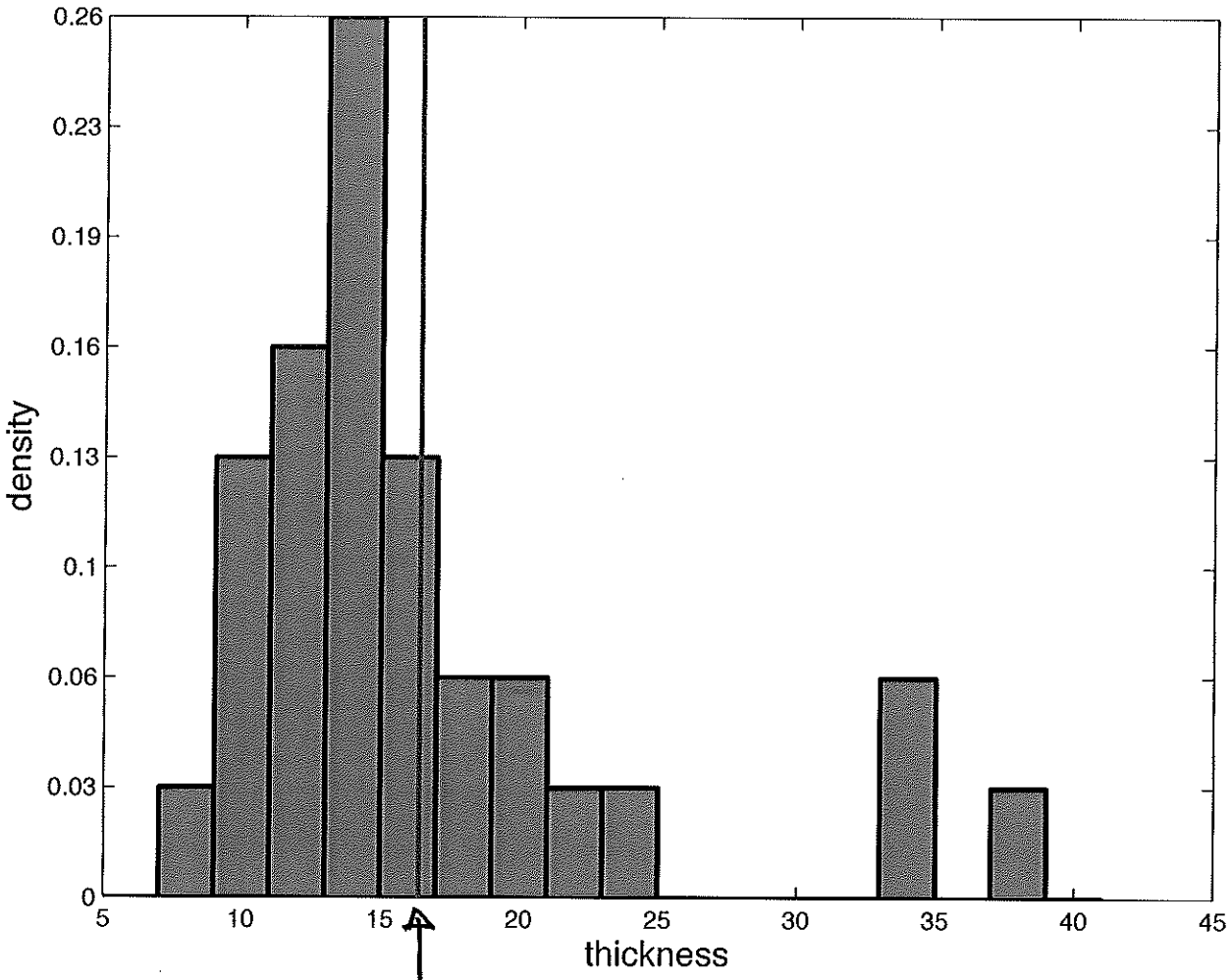
31 different species were measured, and the average leaf thickness computed for each species.



Most leaves have thickness around 14 with a spread of ~4.

outliers. In this case there are leaves that are much thicker than most.

these are <u>not</u> errors (as the values < 10 for the dice data were).



mean. — appears to be a bit too high.

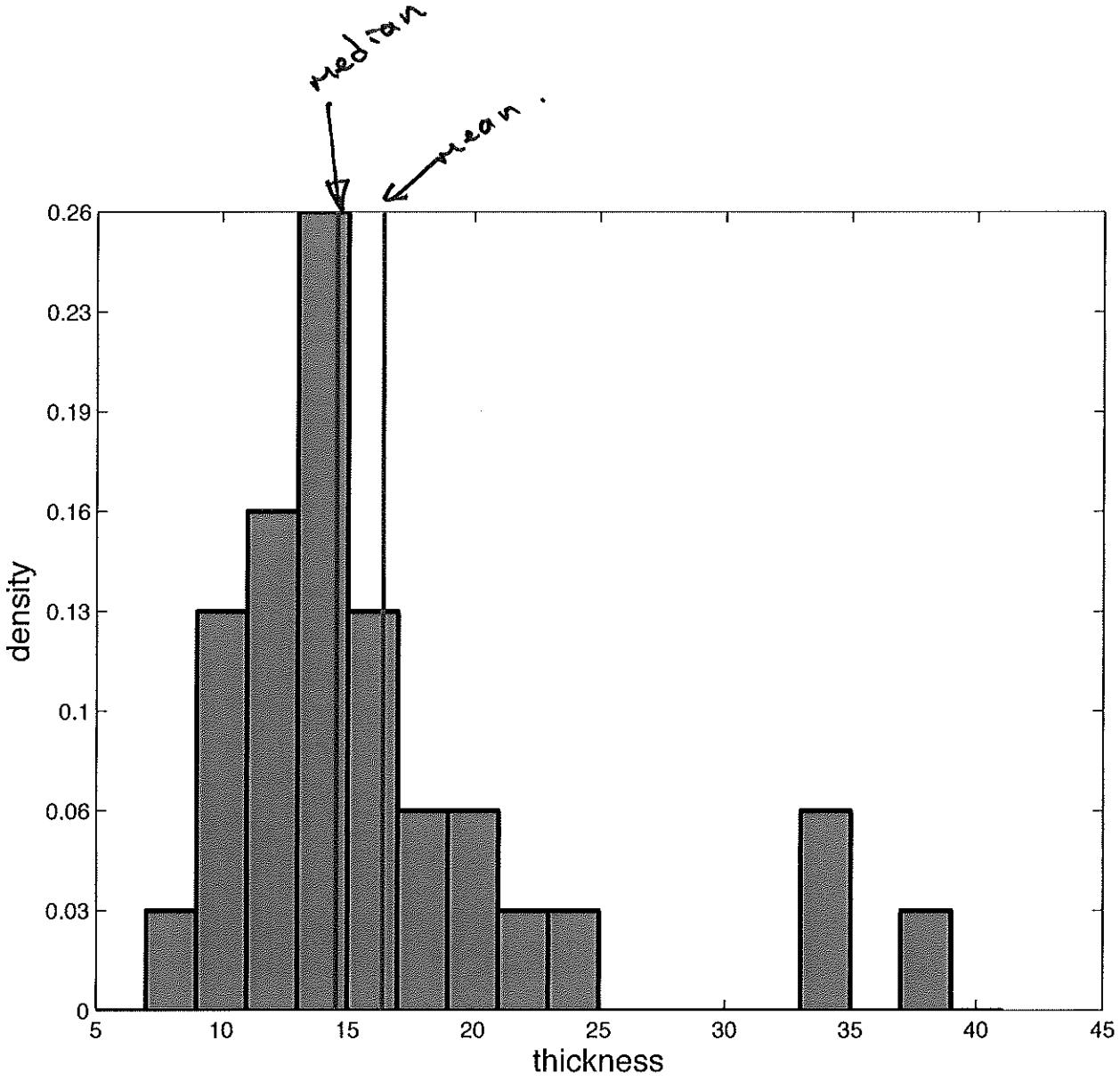
The mean is sensitive to the presence of outliers.

Average of the 31 measurements is 16.4

Is there a different measure of the center that is not as affected by the presence of outliers?

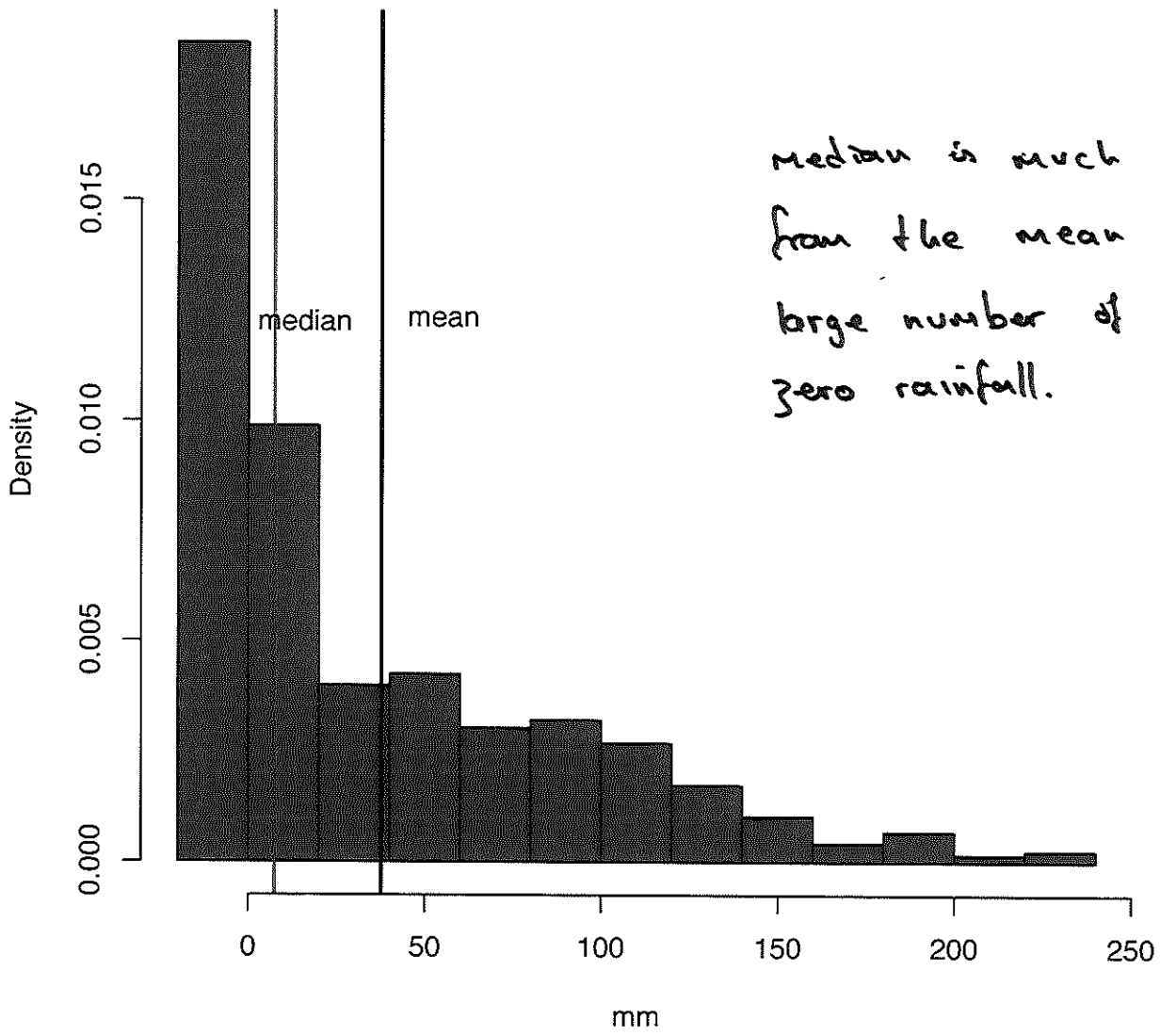
The median is often a better measure in the presence of outliers, or when the distribution is not symmetric.

Median: half of the area of the histogram is to the left of the median, and half is to the right



median is more in the "center" of the distribution.

Because there are only a few outliers, the difference between mean + median is not very large.



median is much different from the mean due to the large number of days with zero rainfall.

Computing the median.

Sort the data in ascending order

leaf thickness.

14 17 **17** 18 19

if there's an odd number of values, the median is the middle one.

33 36 36 37 39 40

$\frac{36+37}{2} = 36.5$

if there's an even number of values, the median is the average (mean) of the two middle numbers.

[Be clear of the difference between the definition of the median, and this particular method for finding its value]

Example.

A business owner says that the average salary of the 10 people in her business is $110,000.

the workers all claim to be being paid minimum wage.

what's going on?

9 workers are paid $10k.

the owner makes $1M.

"average" (mean) is $109,000

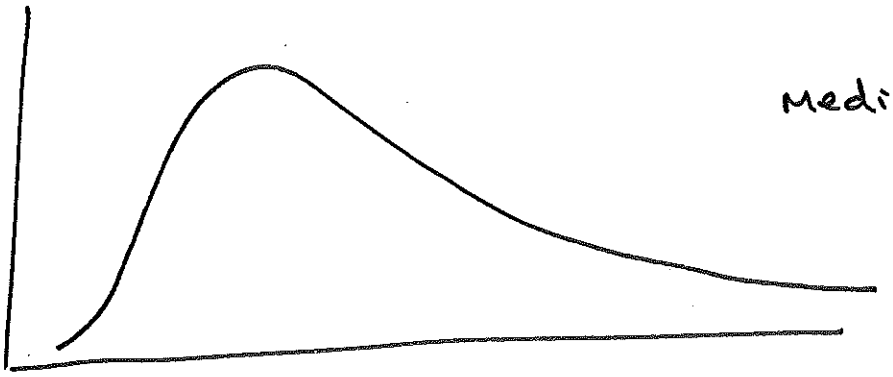
median is $10,000.

Relationship between mean + median

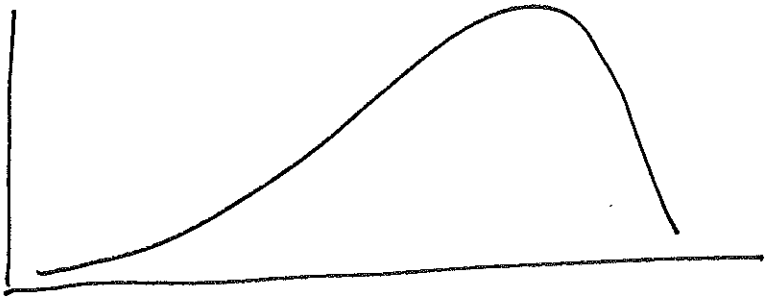
tells you about the shape of the distribution.

symmetric distribution / histogram

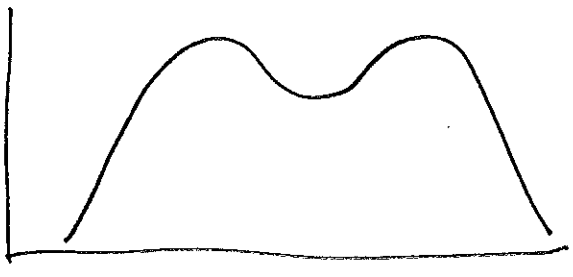
mean = median.



median < mean.



mean < median.



bimodal distribution.

- can't be summarized so easily.