measures of averages and variation

- lies, damn lies and ...
  mode(s), median and mean
- square people
  variance and standard deviation

average?

three typical measures:

- mode(s):

  “more people use dogo than any other dog food”

- median

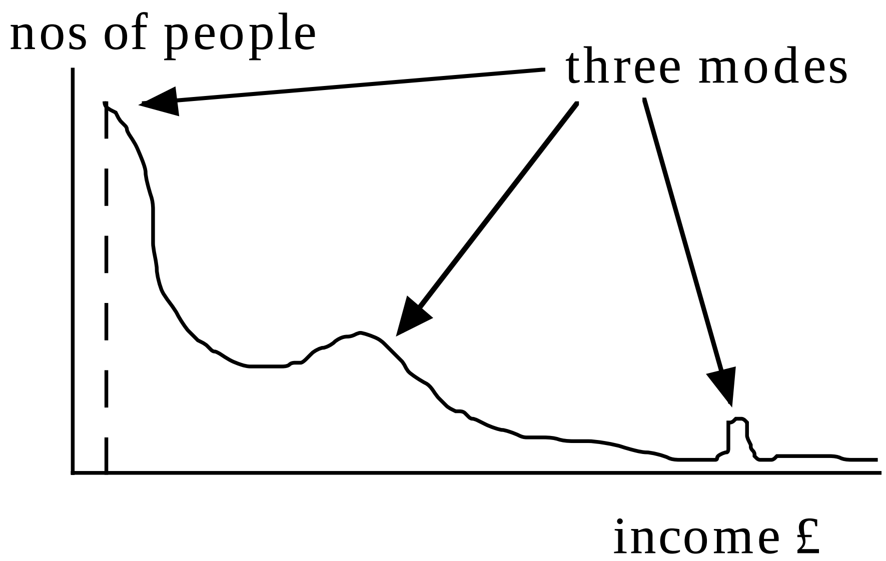
  “half of all salaries are greater than £15000 p.a.”

- mean

  “if salaries were divided evenly . . .”

mode(s)

- not widely used
- may have more than one mode
- the bump may be anywhere!
- sensitive



sensitivity of mean

- one big value ...
- union quotes median
- employer the mean
- lies, damn lies ...

J. Bloggs	3500
F. Mole	5600
K. Giles	8000
J. Smith	8300
B. Roberts	8450
S. Claus	8450
A. Jones	8680
H. Lee	15750
M. Warren	17500
T. Smyth-Boule	200000
	28423

median salary £8450

mean salary!

why use the mean?

- median is more robust
- mean is more manipulable

	number of people	mean salary	median salary
group 1	10	15000	12500
group 2	10	23000	16000
grp 1 & grp 2		19000	?

measures of variation

difference from mean

square of difference

2	10	100
7	5	25
8	4	16
10	2	4
11	1	1
12	0	0
12	0	0
13	1	1
13	1	1
15	3	9
18	6	36
23	11	121
	4	26

average difference

variance

2
7
8
10
11
12
12
13
13
15
18
23
12

inter-quartile range = 14–9

mean

standard deviation

$\sigma = \sqrt{\text{variance}}$

which is best?

a bit like averages . . .

- inter-quartile range is robust
- variances add up
- standard deviations meaningful

square people

if data is people buying ‘dogo’

variance is 26 square people!

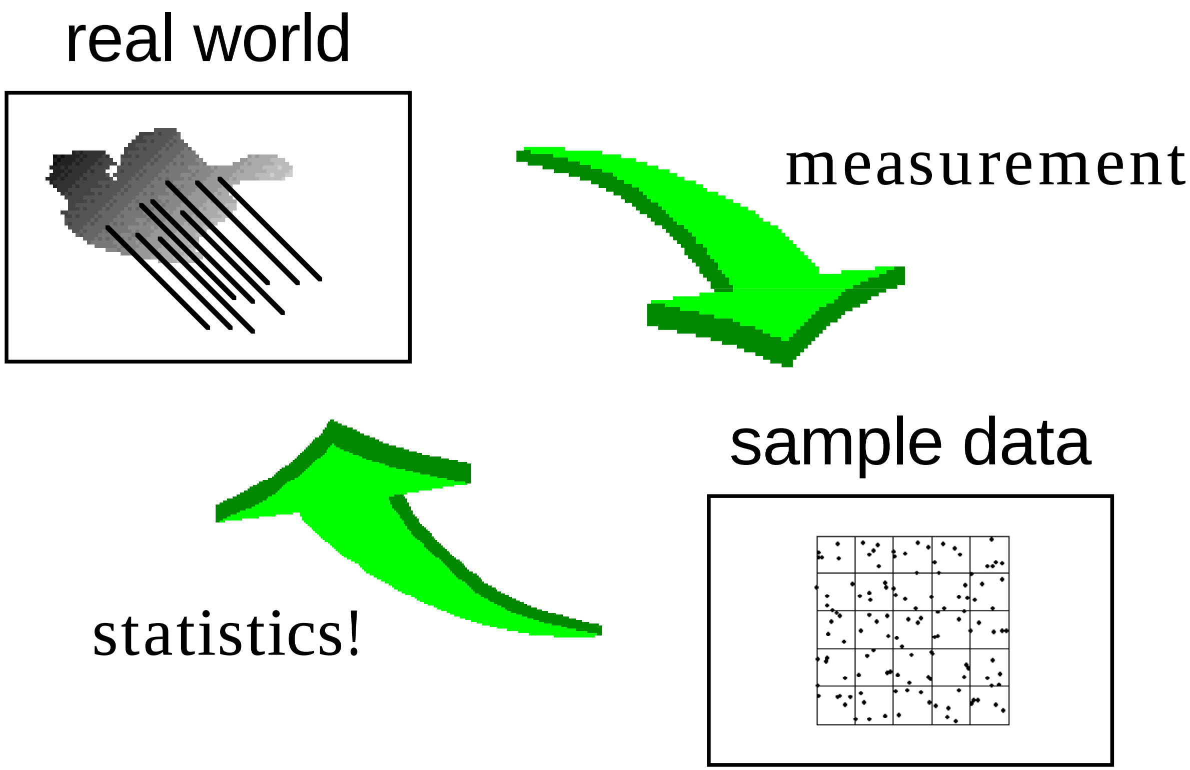
standard deviation

$$\sigma = \sqrt{\text{variance}}$$
$$= 5.1 \text{ people}$$

the ‘real’ world

- the sample – actual measured data
- the population
  - large set from which the data is drawn
  - especially for surveys etc.
- the ideal
  - the ‘typical’ user, the fair coin
  - unrepeatable events – the fall of a raindrop
  - a theoretical distribution

the job of statistics



different means

① average of the measured data

~ sample mean

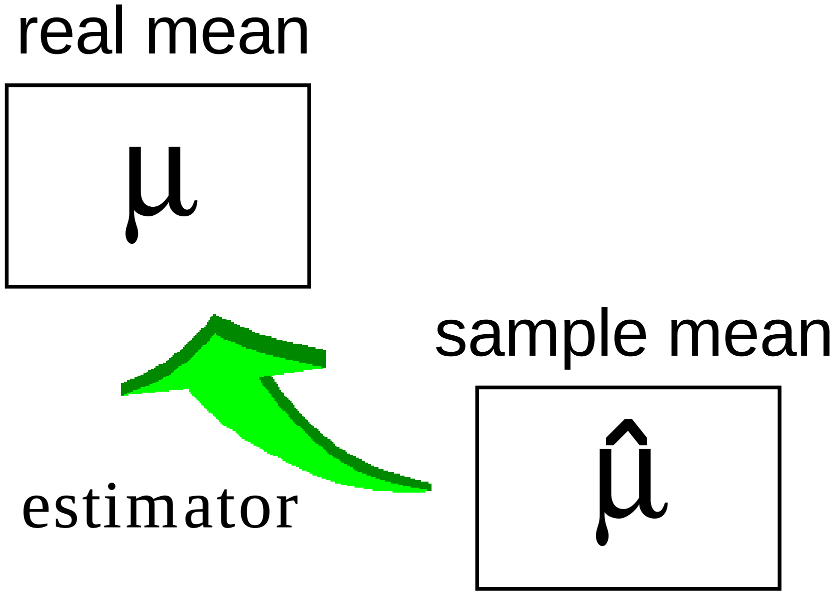
② average of the ‘real’ world

~ population mean

③ theoretical mean of the ‘distribution’

e.g. mean die score = 3.5

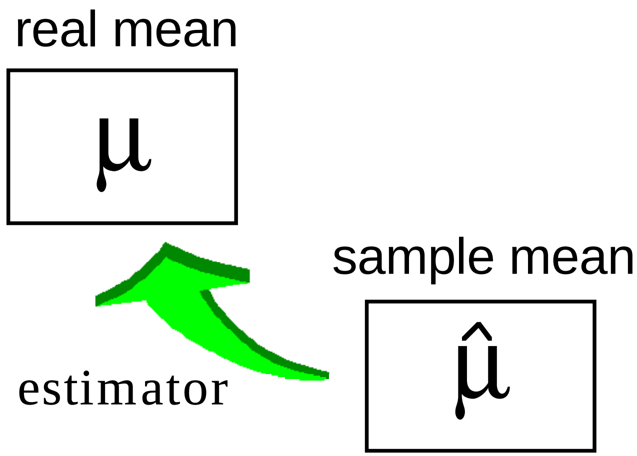
estimating the mean



sample mean estimates
real (population) mean

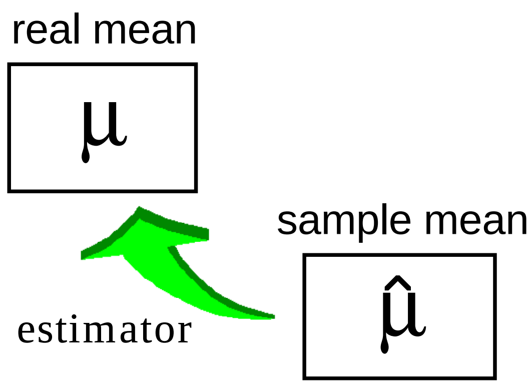
strange but true

the mean
of the mean
is the mean



i.e. theoretical mean
of sample mean
is real mean!!!!!

law of large numbers



<u>if</u> samples are independent

(or nearly so)

bigger sample $\Rightarrow$ better estimate

how good an estimate

- each data item has some variability

  head=1 / tail=0: 0 0 0 1 1 1 0 1 1 1 0 1 1 1 0 0 1 0 1 1

- sums of data items have variability

  nos of heads: 12 11 9 13 8 8 8 11 8 11

- means of data item have variability

  averages: 0.6 0.65 0.45 0.65 0.4 0.4 0.4 0.55 0.4 0.55

better = less variability

variability of sums

variances add up*:

variance(sum of 100 items)

$= 100 \times$ variance(each item)

standard deviation $= \sqrt{\text{variance}}$

s.d. of sum of 100 items

$= 10 \times$ s.d. each item

square root rule: σ(n items) $= \sqrt{n}\ \sigma$(each item)

i.e. bigger, but proportionately less

* only if items are independent (actually closely related to Pythagoras' theorem!)

variability of mean

mean is sum / nos. of items:

$$\sigma(\text{mean of 100 items}) = \sigma(\text{sum each item}) / 100 = \sigma(\text{each item}) \ / \ 10$$

square root rule for means:

$$\sigma(\text{mean of n items})^* = \frac{1}{\sqrt{n}}\,\sigma(\text{each item})$$

* called standard error (s.e.) of mean

so what?

experiments, data collection etc....

to halve the variation
need 4 times as many subjects

solved it?

① seeing through randomness

use sample mean as estimator

② knowing when you have

$$\sigma(\text{mean}) = \sigma(\text{item}) / \sqrt{n}$$

? what is σ(item)

estimate it from sample!

estimating σ(item)

use sample variance / s.d.
as estimate
of real variance

real variance

$$\boxed{\sigma^2}$$

estimator

sample data

$$\boxed{\frac{\sum(x-\hat{\mu})^2}{n-1}}$$

N.B. only an <u>estimate</u>

OK . . . but a tid bit small on average
(biased estimator)

☆ that's why stats. formulae are full of $\sqrt{n-1}$

in short ...

- estimate value using sample mean
- accuracy of mean ~ $\dfrac{1}{\sqrt{\text{nos in sample}}}$
- estimate accuracy of sample mean ...
  ... using variation within sample

drunkard's walk

- a drunk wanders home

→ sometimes he takes one step forwards

sometimes one step back ←

? after n steps
how far is he from where he started

! another example of $\sqrt{n}$ behaviour