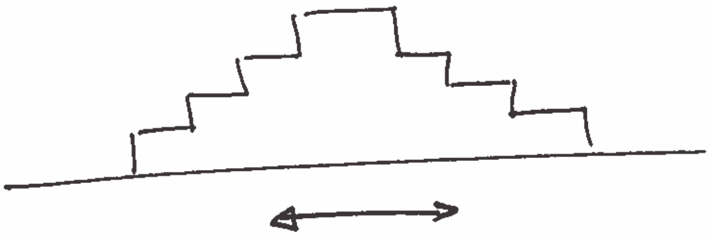
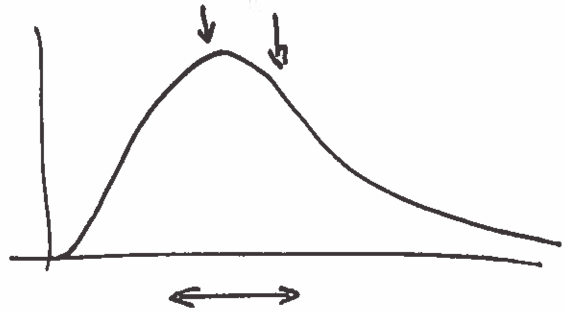
Variability

Randomness

Uncertainity





Quantify
the variability
in the data

<u>Range</u> : max data value − min data value.

− not <u>robust</u> in the presence of outliers.

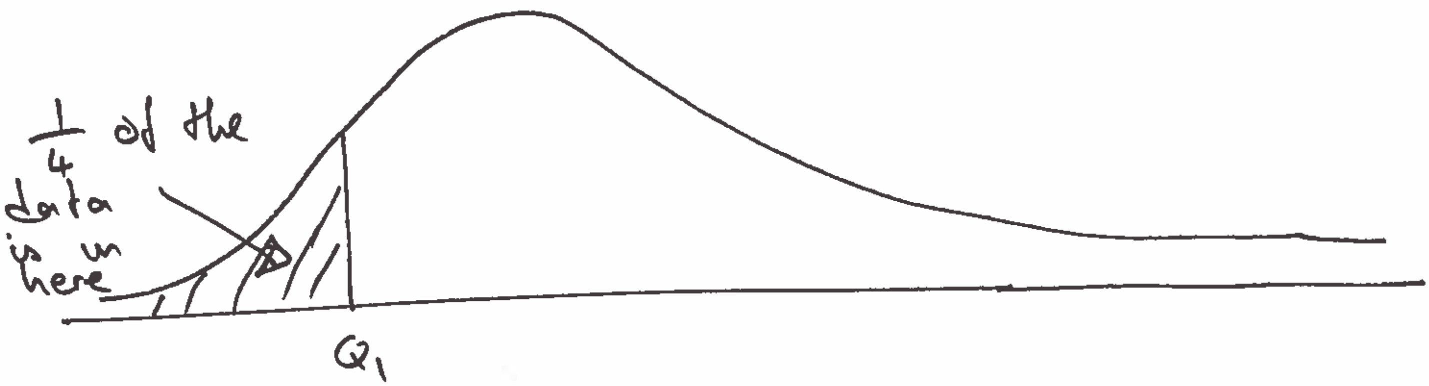
⇒ get rid of the outliers !

compute <u>quartiles</u>.

1st Quartile : one quarter of the data is less than Q_1

2nd Quartile : median

3rd Quartile : $3/4$ of the data is less than Q_3



↳ Q_1 is a <u>data</u> value

A measure of the spread of the data is Inter-quartile range

$$Q_3 - Q_1$$

→ robust to outliers

→ good for non-symmetric distributions

Quartiles are just specific percentiles.

- given a percentile, eg 10^{th}-percentile. count up through the sorted data until you have seen 10% of the data. The data value you have reached is the 10^{th} percentile value.

- given a data value, its percentile is given by the fraction of the data that has smaller values.

Another measure of the spread, (that is more convenient mathematically) is the <u>standard deviation</u>.

What is the "average size" of a set of numbers?

leaf thickness 33 36 36 37 39 40

"size" is around 36 or 37

what about: 0 5 -8 7 -3

average is <u>0.2</u>.

most of the numbers are quite different from 0.2.

→ need to make all the numbers positive.

1/ square all the numbers 0 25 64 49 9.

2/ compute average of the squared values $\dfrac{0+25+64+49+9}{5} = 29.4.$

3/ take the square root $\sqrt{29.4} = \underline{\underline{5.4}}$

this is called the Root Mean Square value

(RMS)

(has nicer mathematical properties than just dropping the negative signs)

Measure of the <u>spread.</u>

- RMS deviation about the mean

  - on average, how far away are the data from the mean.

<u>Computing the SD.</u>

20, 10, 15, 15 $\leftarrow$ data.

1/ compute the mean $\frac{20+10+15+15}{4} = \frac{60}{4} = 15$

2/ compute the deviation of each data value from the mean.

ie the difference between each data point and the mean.

20 − 15	5
10 − 15	−5
15 − 15	0
15 − 15	0

3/ Calculate the RMS of the deviations

$$\sqrt{\frac{(5)^2 + (-5)^2 + 0^2 + 0^2}{4}}$$

$$= \sqrt{\frac{25 + 25 + 0 + 0}{4}} = \cancel{\sqrt{\frac{100}{4}}}$$

$$= \cancel{\sqrt{25}}$$

$$= \sqrt{\frac{50}{4}} = \sqrt{12.5} \doteq \underline{\underline{3.5}}$$

$\uparrow$

S.D.

a measure of the size of the spread about the mean.

Note: Make sure you're using the correct SD button on your calculator

- we're dividing by N, not (N-1)

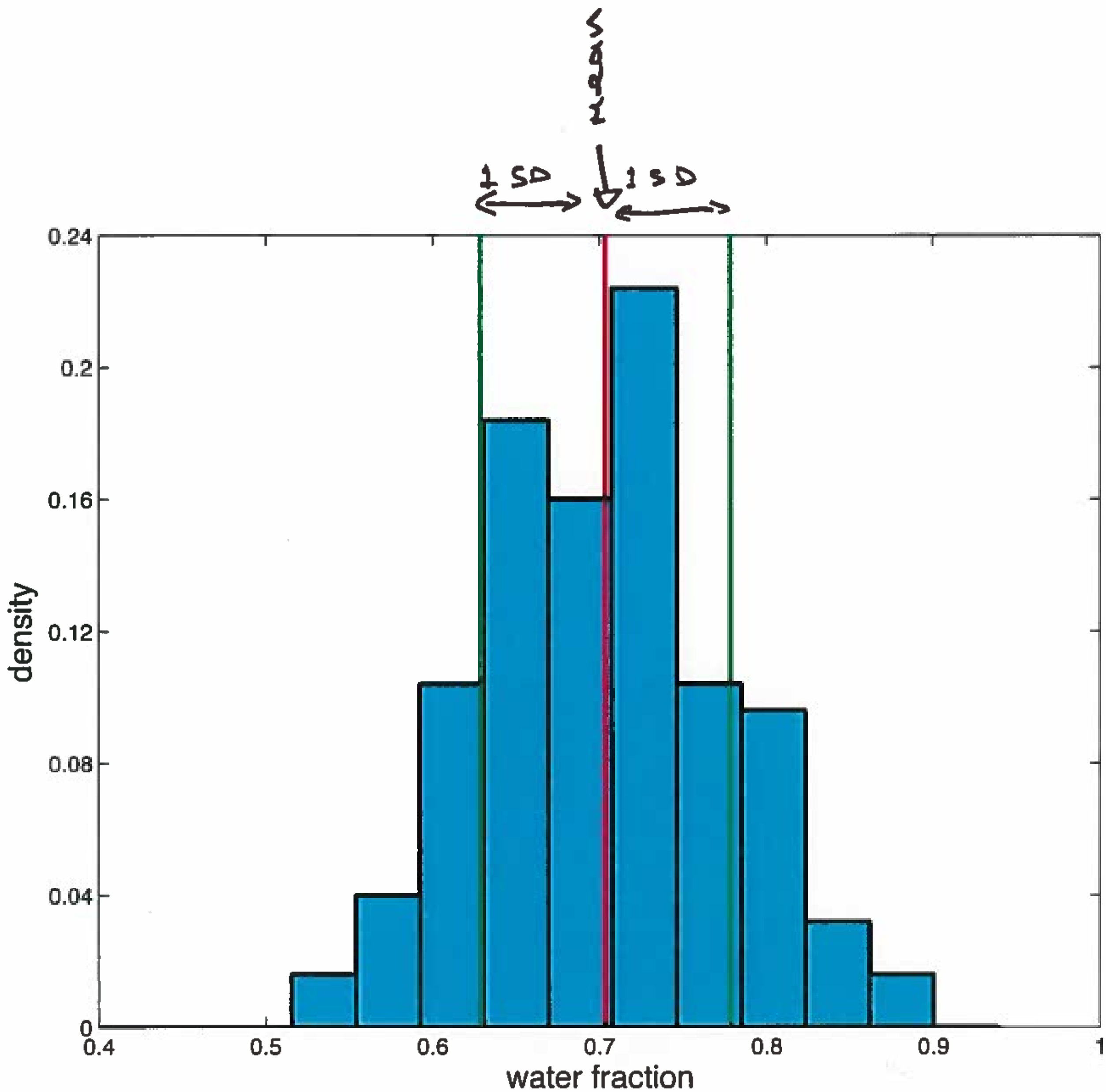
Most observations are within ± 1 SD of the mean.

Few are more than 2 SD away from the mean.

68% of entries (≈ 2 in 3) are within 1 SD of mean.

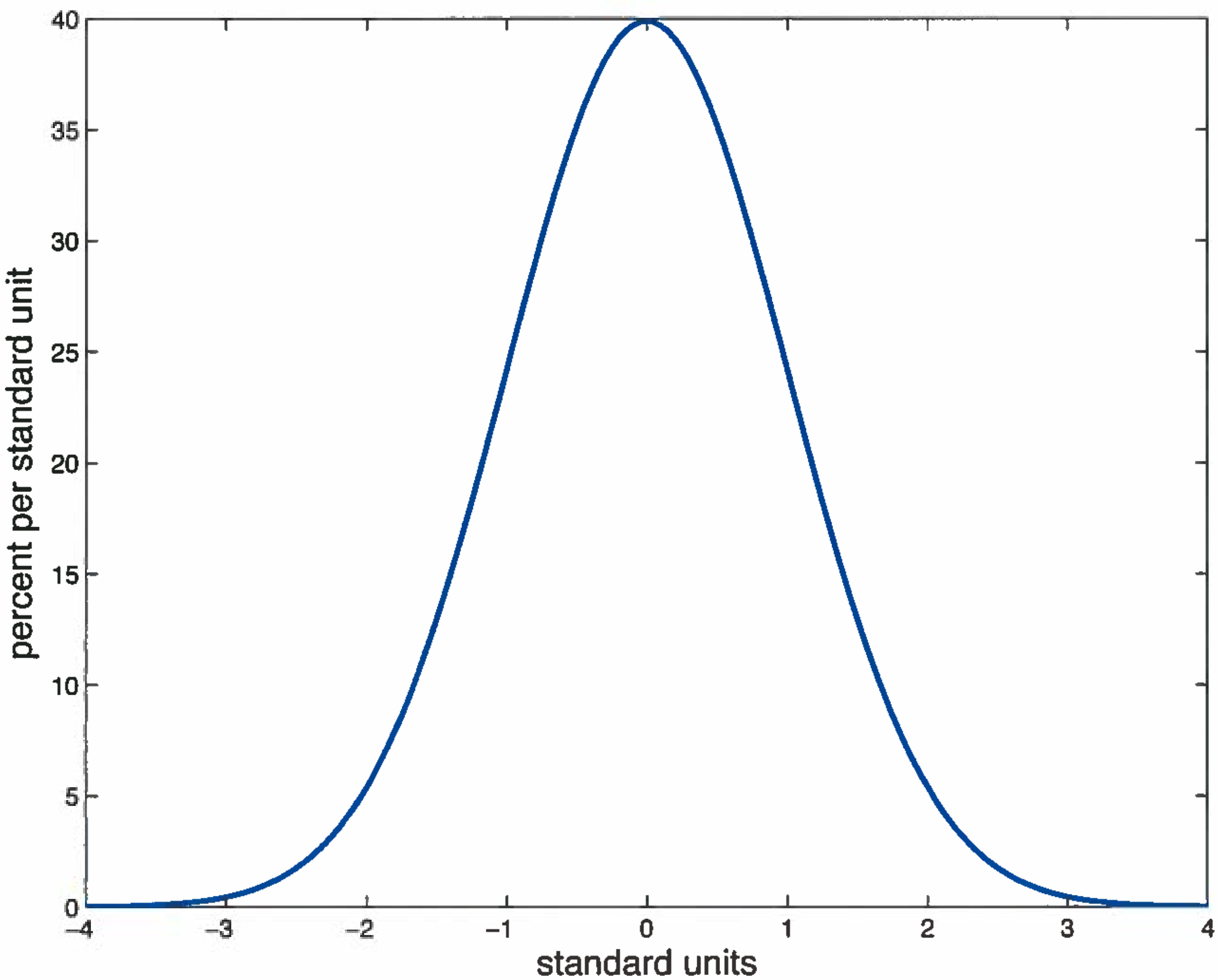
95% (19 in 20) 2 SD

this holds best for ~~symmetric~~ data with a symmetric distribution, but is often ok even for data with non-symmetric distributions.



About 68% of the data is within 1 SD of the mean.

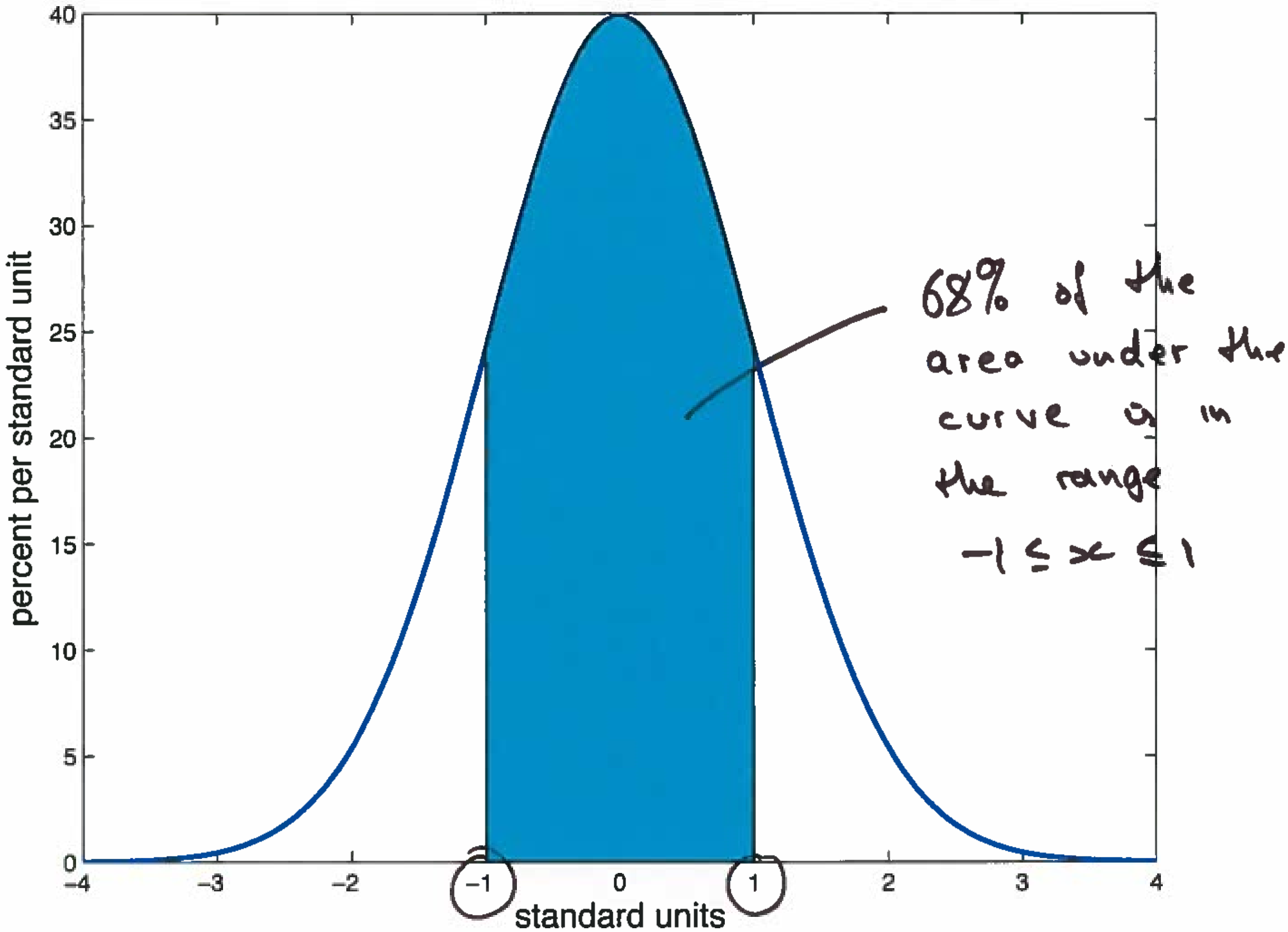
Standard Normal curve



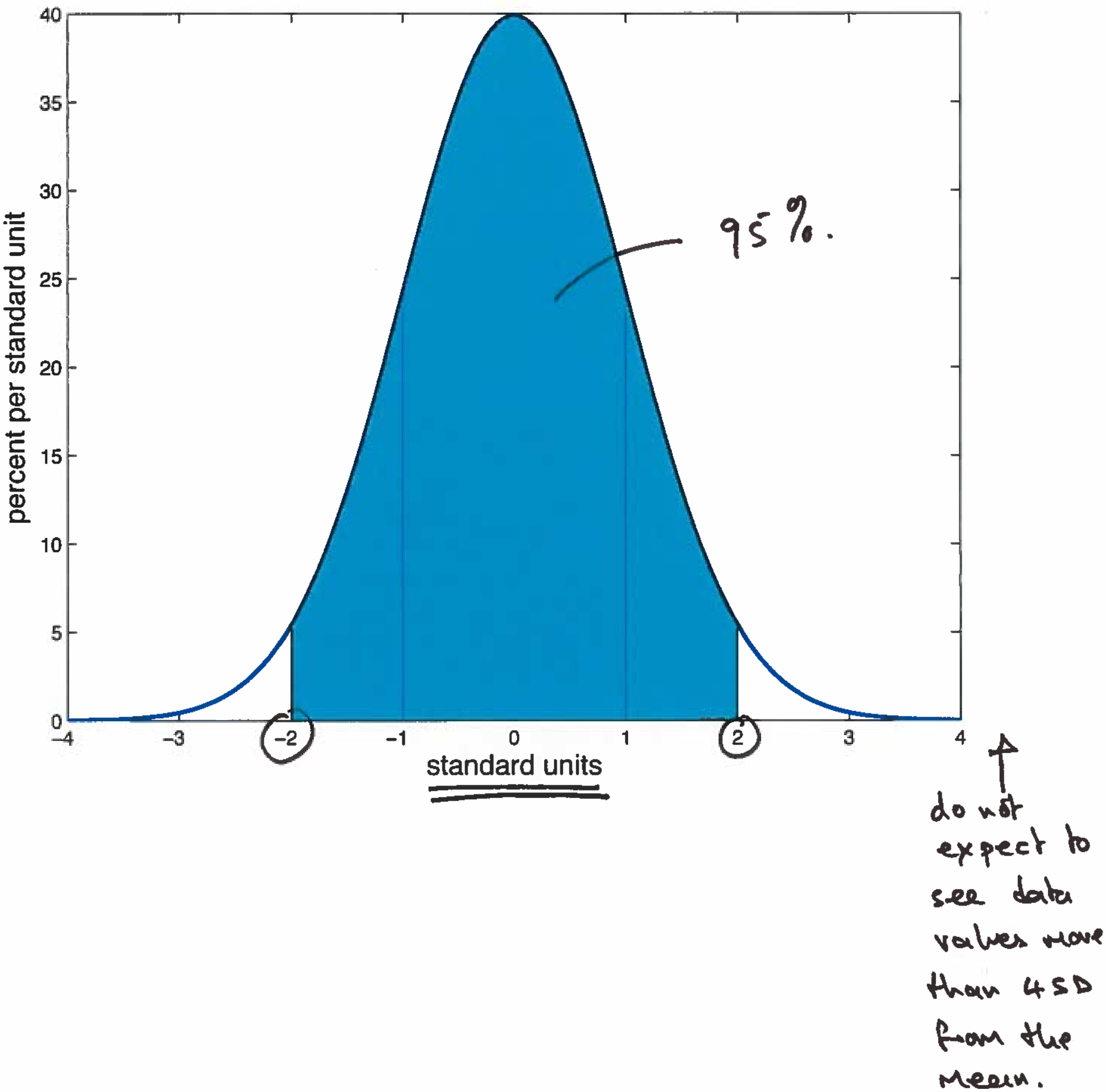
Mean = 0

Median = 0 (symmetric)

SD = 1.



percent per standard unit
standard units
68% of the area under the curve is in the range $-1 \le x \le 1$


percent per standard unit
40
35
30
25
20
15
10
5
0
−4
−3
−2
−1
0
1
2
3
4
standard units
95%.
do not expect to see data values more than 4 SD from the mean.

Normal Approximation for Data.

Very many (but by no means all) data histograms can be approximated by the Normal Curve.

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2}$$

For Normal curve.

between	-1 and $+1$,	area under curve	$\doteq$ 68%
	-2	$+2$	95%
	-3	$+3$	99.7%

x - axis is in standard units - what does that mean?

standard units are
the number of SD
away from the mean.

Example

Women in a survey had mean height 63.5 in

SD 3 in

One particular woman was 69.5 in tall.

convert to standard units:

69.5 is 6 in more than average

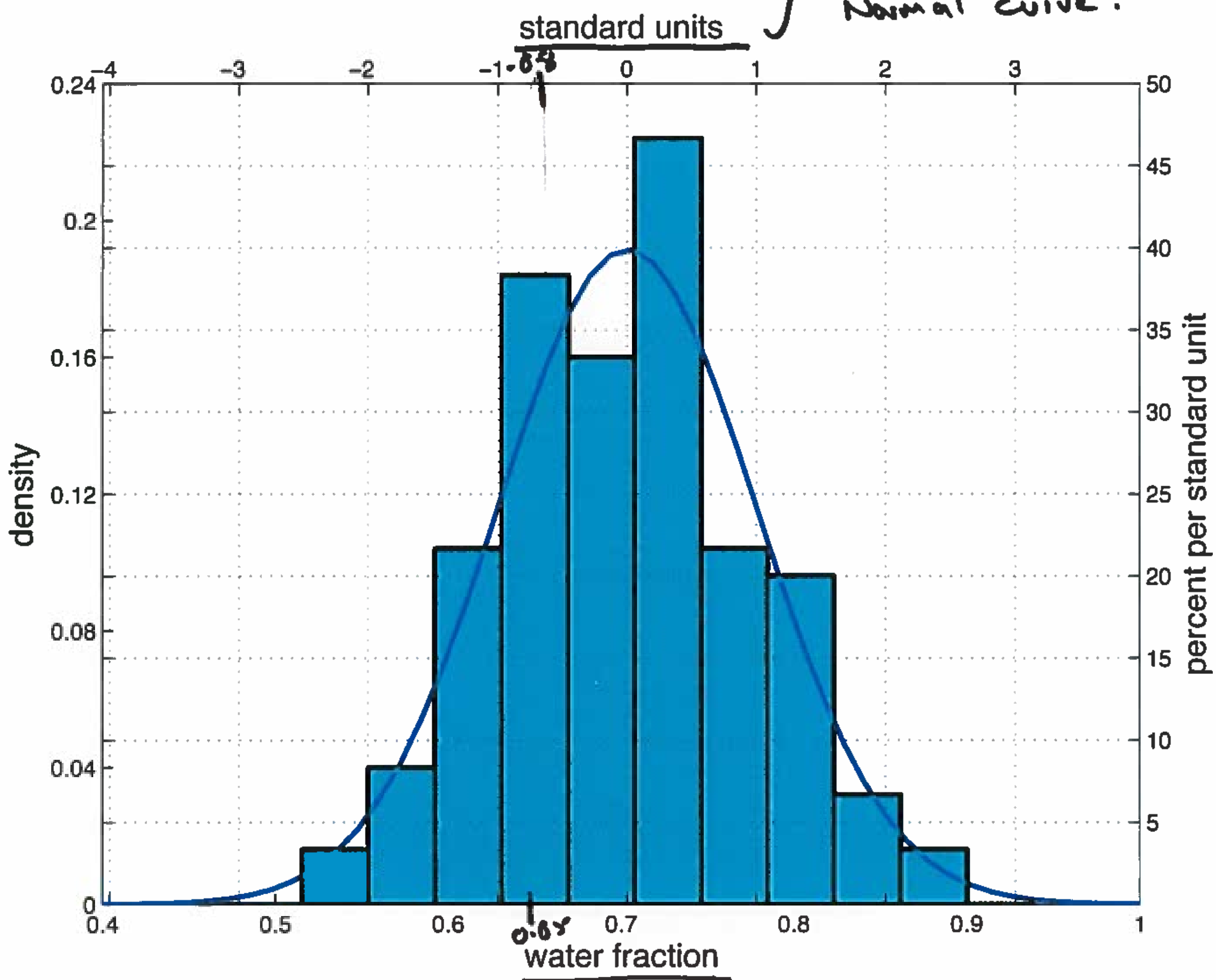
6 in is twice SD

so 69.5 in is 2 in standard units.

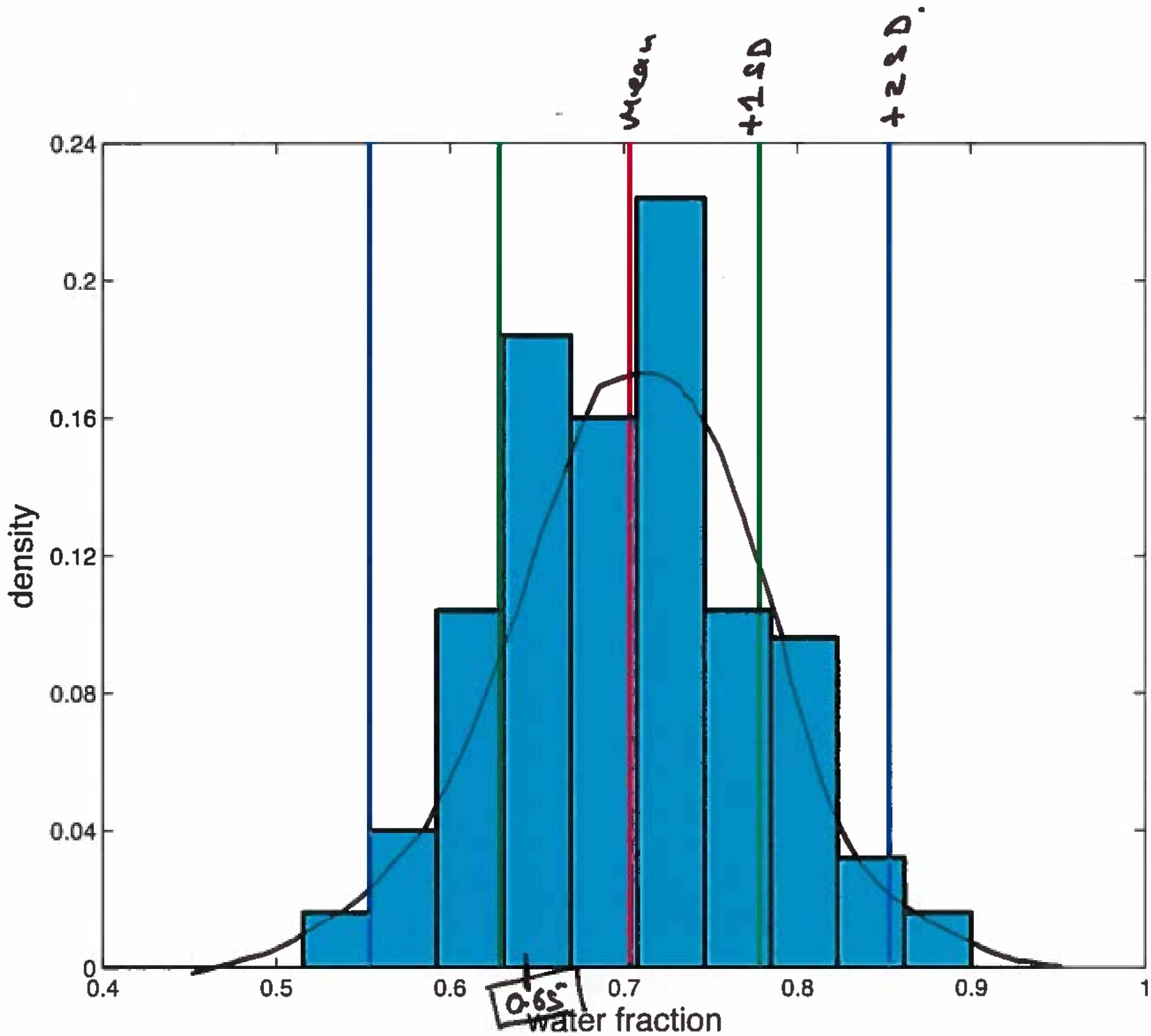
Another woman was 62 in tall.

convert to standard units $\dfrac{62 - 63.5}{3} = \dfrac{-1.5}{3} = -0.5$

units for the
Normal curve.



units for the
histogram.



→ ~ 68% of data are within 1 SD of mean

~ 95% 2

Converting into standard units allows us to use a *single* Normal curve.

To approximate the % of entries in an interval in *data scale*

(eg % of leaves with water fraction < 0.65)

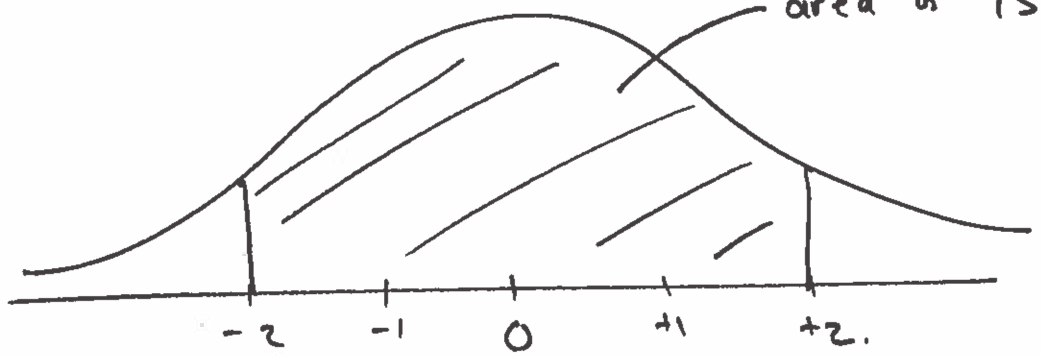
→ convert that interval into standard units.

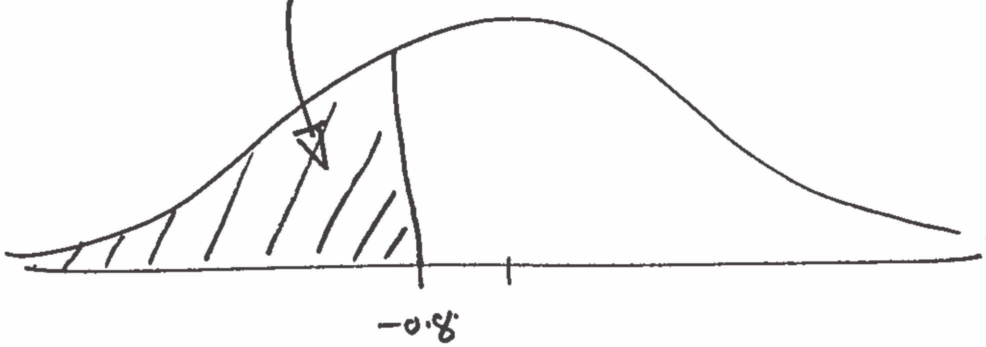
→ find the area under the normal curve in this interval.

(area under normal curve to left of ≈ -0.8).

- We find areas under the normal curve by using the table in the back of the textbook.
  - the table has areas for *symmetric* intervals.

area is 95.45%.





this is the area in the table.

= 58%

Two red areas together are 100 - 58 = <u>42%</u>.

By symmetry One of the red areas is $\frac{42}{2}$ = 21%.

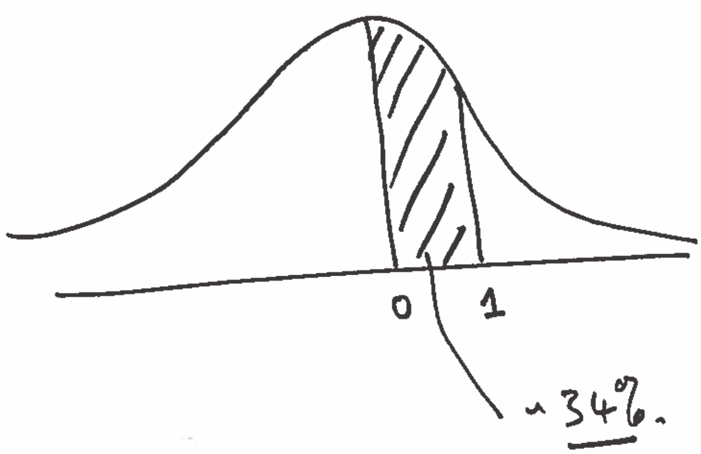
By combining

- central area from the table

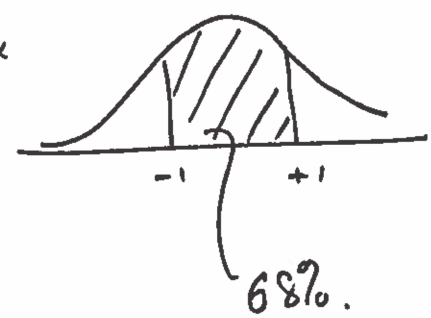
with knowledge that

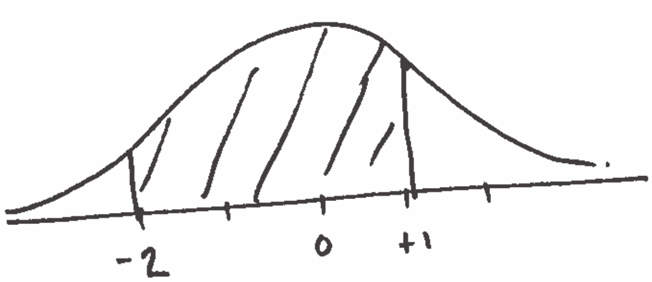
- the curve is symmetric
- the total area is 100%

we can compute the area underneath the curve for <u>any</u> interval.



$= \frac{1}{2} \times$





$=$

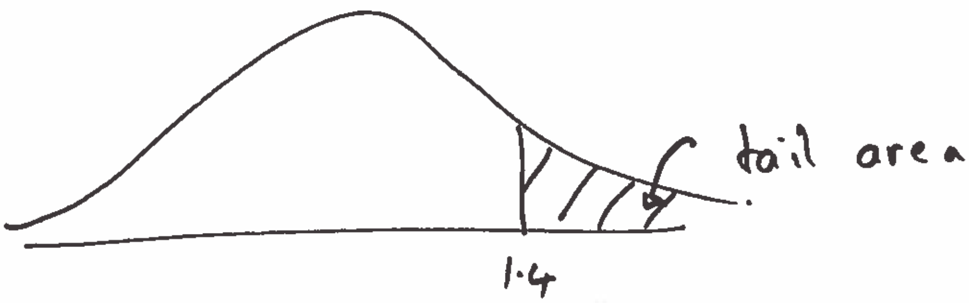
$+$

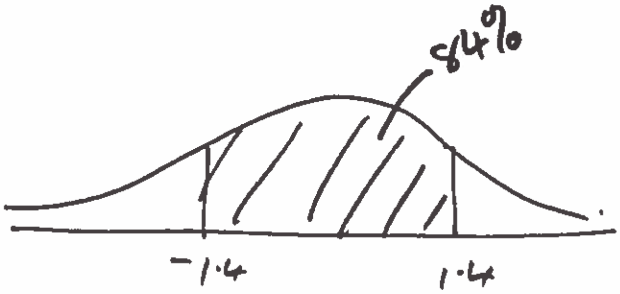
$= \frac{1}{2}$

$+ \frac{1}{2}$

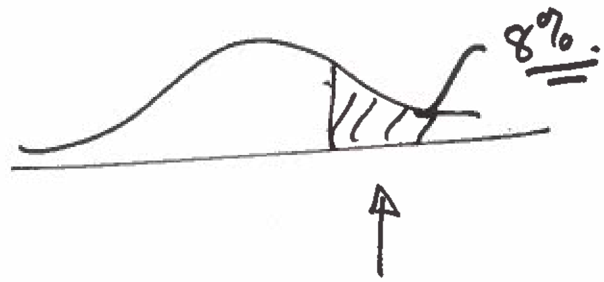
$$= \frac{1}{2} \times 95 + \frac{1}{2} \times 68\%$$

$$= \underline{82\%}.$$





together have
100 - 84 = 16%.



ie we would expect about 8% of our data set to have values more than 1.4 SD above the mean

Normal Approximation for Data.

Key: convert the region of data space that we're interested in into standard units, using the mean + SD of the data to do so.

Use the standard normal curve to obtain the % of the data that falls into the region.

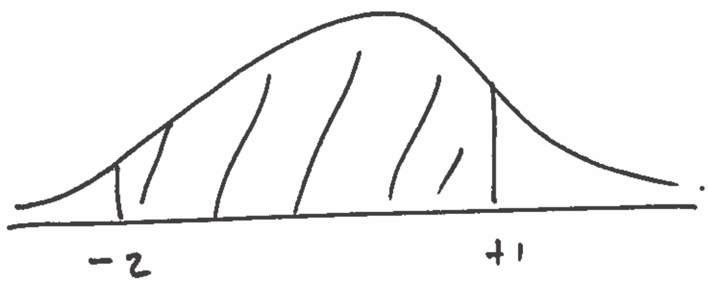
Example. the heights of a group of men had mean 69 in SD 3 in

What % of the men had heights between 63 and 72 in?

Assuming that the histogram of the heights follows the Normal curve.

data space
(heights in inches)

standard
units.



from before, this is

$$\frac{95}{2} + \frac{68}{2} \simeq \underline{\underline{82\%}}$$

$\frac{68}{2}$.

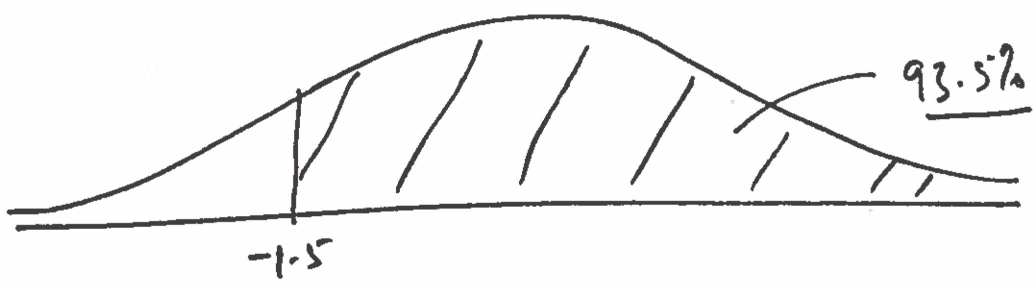
$\frac{95}{2}$.

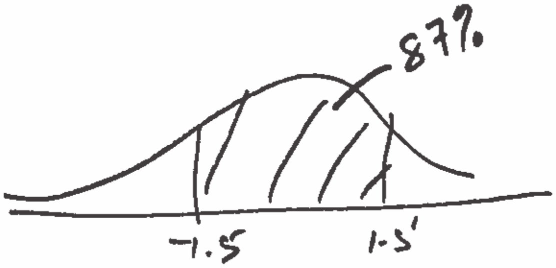
womens heights   mean 63.5

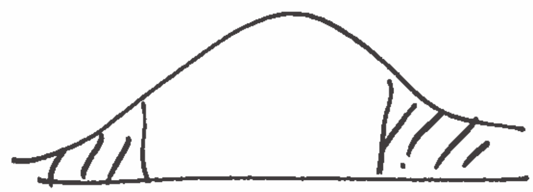
SD 3 in.

% with heights > 59 in

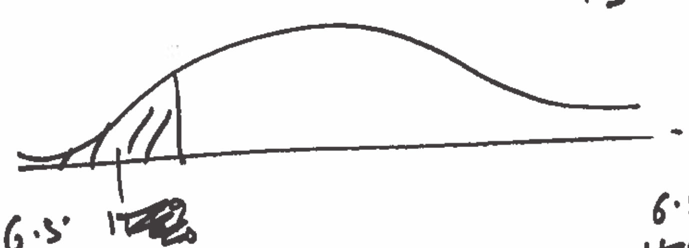
59 in standard units $\frac{59 - 63.5}{3} = -1.5$







together have 13%.



100 - 6.5

= 93.5%