A few comments about tests of significance.

1/ Thresholds 5% "statistically significant"
1%. "highly statistically significant"

These are conventional but arbitrary.

Don't over-interpret the difference between
$p = 4.9\%$ and $p = 5.1\%$.

2/ If you do enough tests, eventually you will get a statistically significant result (even if nothing is actually going on)

- chocolate + weight loss study

Deciding your data analysis methodology (which observations to exclude as outliers, how to divide into sub-groups, etc) after you have collected the data.

3/ One sided vs Two sided tests.

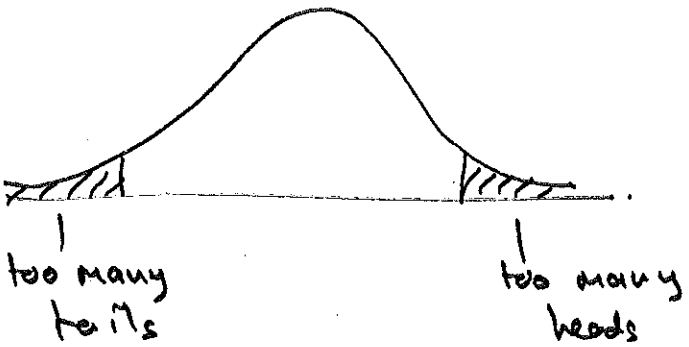
eg is this a fair coin?

does the bottle filling machine over / under fill bottles?

- in both cases, deviations in <u>either</u> direction are evidence against the null hypothesis.

- our data shows a deviation in <u>one</u> direction only, but when determining the p-value, we must include <u>both</u> tails.

$$z = \frac{\text{observed} - \text{expected}}{SE}$$



If fluctuations in <u>either</u> direction cast doubt on H_0, the p-value is the <u>sum</u> of the shaded areas. — two tailed test

one tailed test - if fluctuations in one direction only are important.

4, the difference between statistically significant and important.

$$z = \frac{\text{observed} - \text{expected}}{SE}$$

consider the mean of sample values

$$SE_{mean} = \frac{SD_{population}}{\sqrt{\#\text{samples}}}$$

as #samples increases, SE becomes smaller.

If (observed - expected) does not change as the sample size increases,

z - statistic increases

and eventually we will have $p < 0.05$

As sample size gets large, very small differences can result in rejecting H_0.

Is that small difference important?

example.

Rural test scores were 25

City test scores were 26

SD = 10

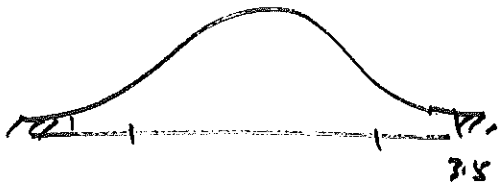
2500 kids in each group.

$$z = \frac{\text{observed difference} - \text{expected diff.}}{\text{SE diff.}}$$

$$SE_{ave} = \frac{10.}{\sqrt{2500}} = 0.2$$

$$SE_{diff} = \sqrt{0.2^2 + 0.2^2} = 0.28.$$

$$z = \frac{1 - 0.}{0.28.} \approx 3.5$$



p. value < 1%.

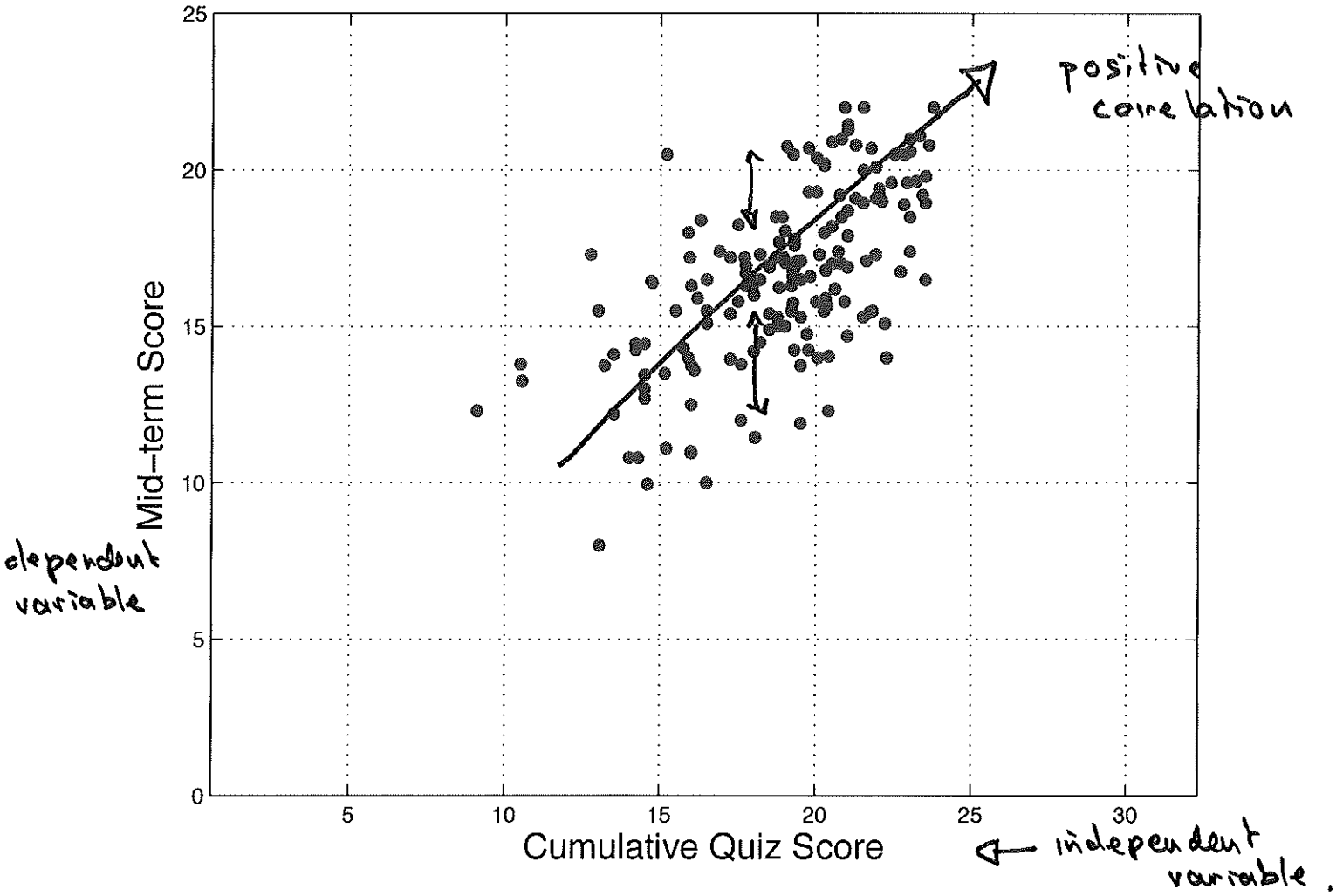
The difference is statistically significant.

But what does a difference of 1 point on this reading test mean

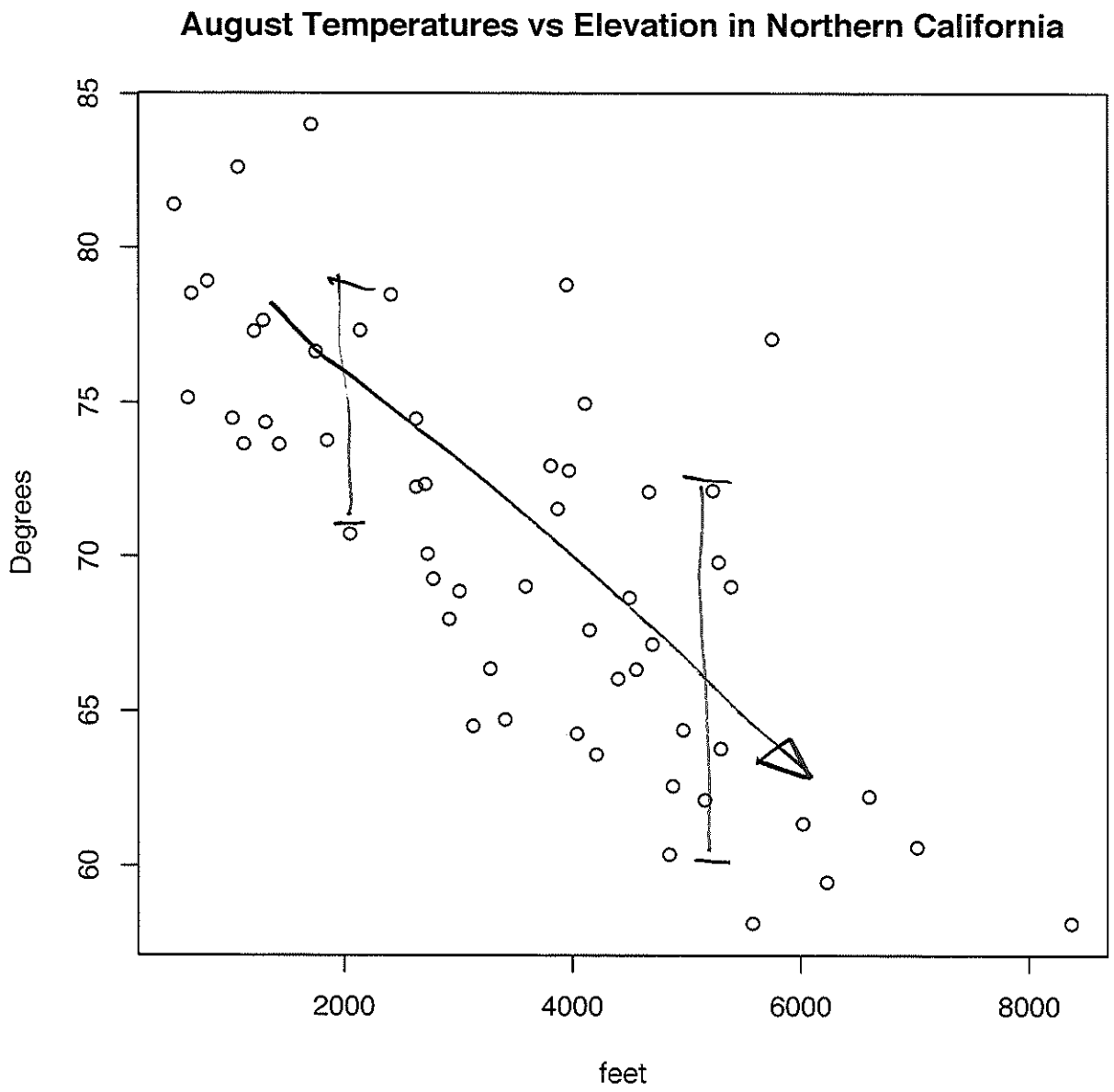
- partial understanding of 1 word out of 40

Q: is this difference important?

5/ Could something else give the effect we've seen?



each (quiz score, mid term score) is plotted as a point



negative correlation

as altitude increases, temp. decreases

Correlation.

How do you study the relationship between two variables?

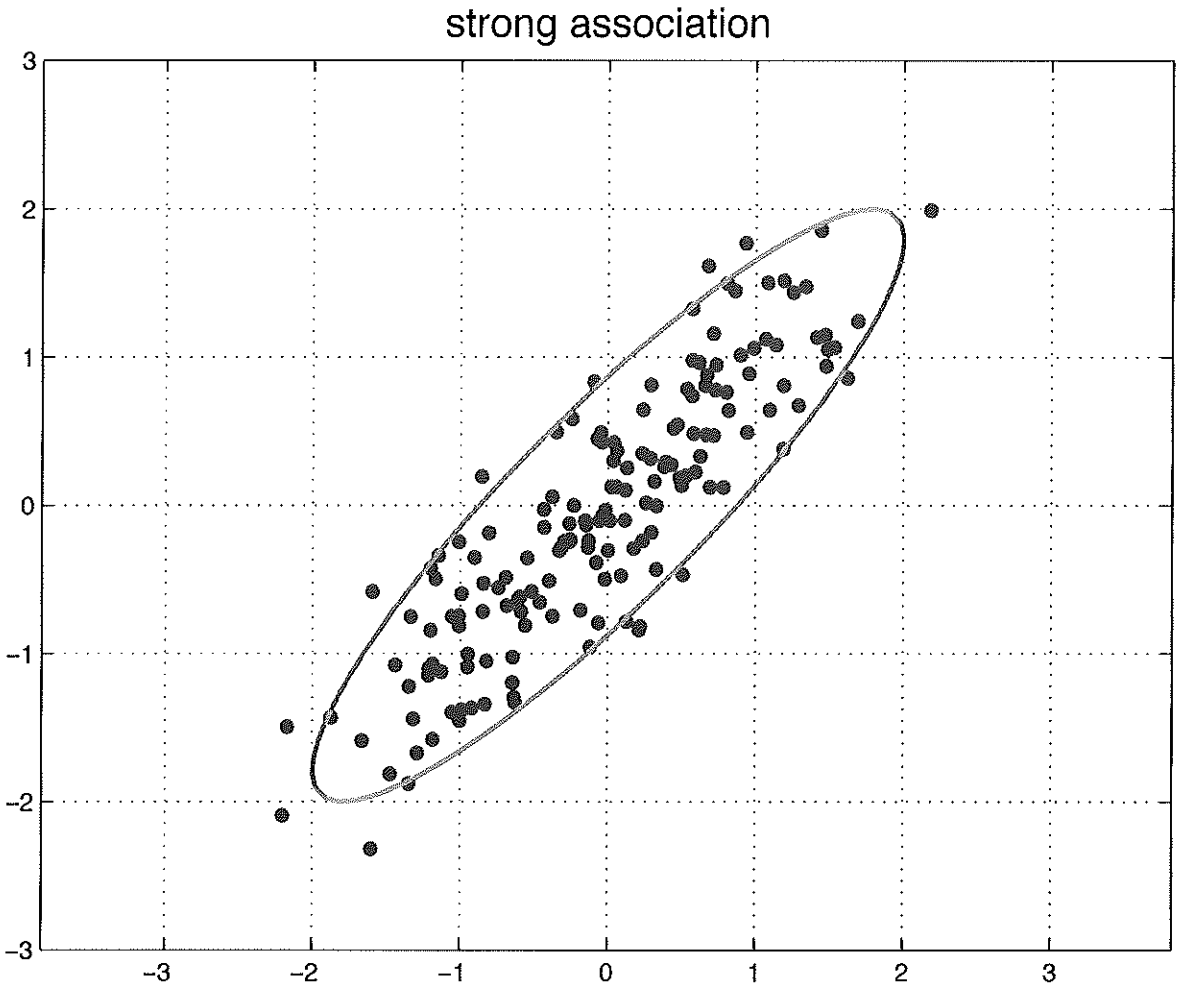
years in school	vs	income
father's height		son's height
quiz scores		midterm score.

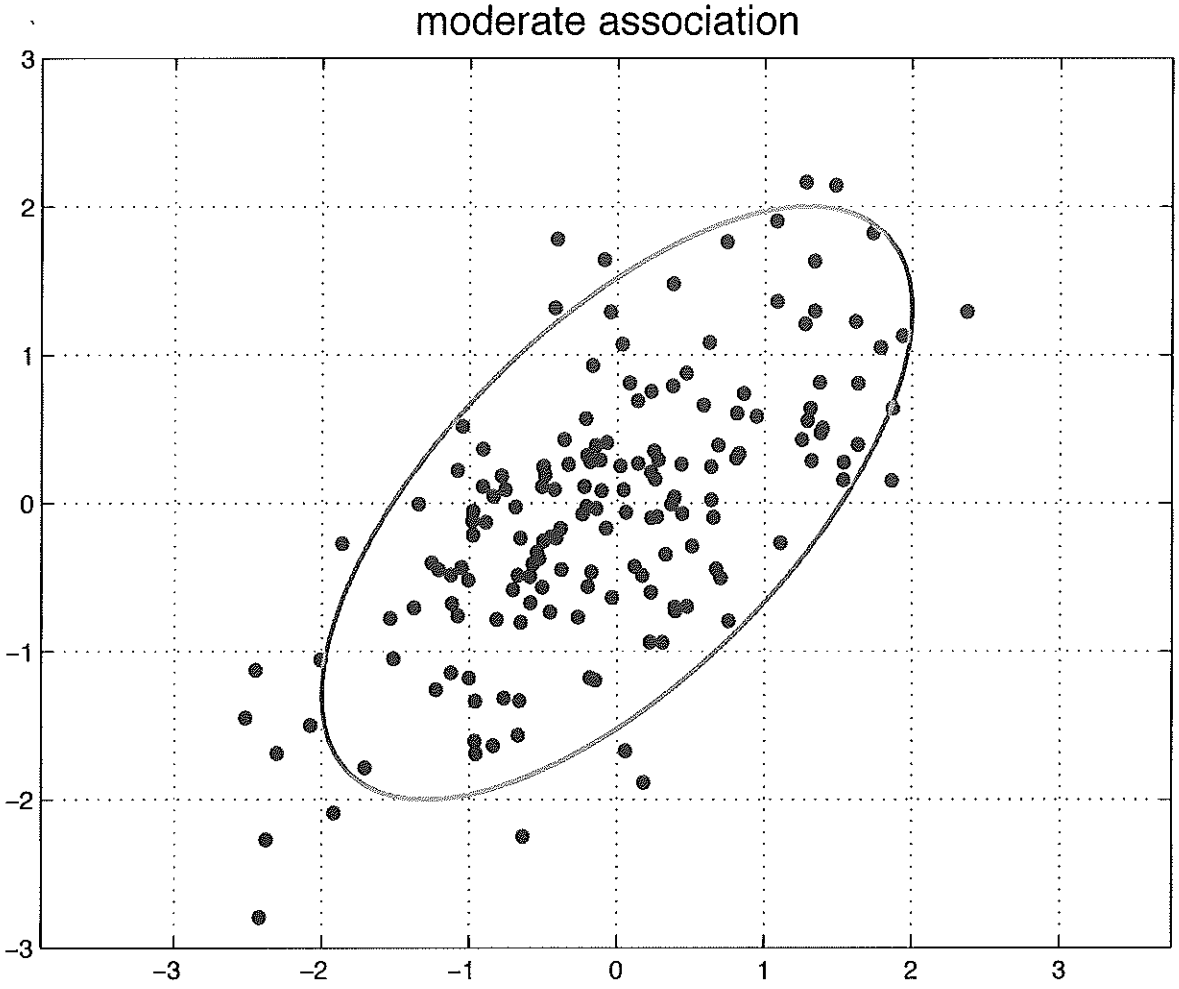
visualize the relation by plotting a scatter diagram

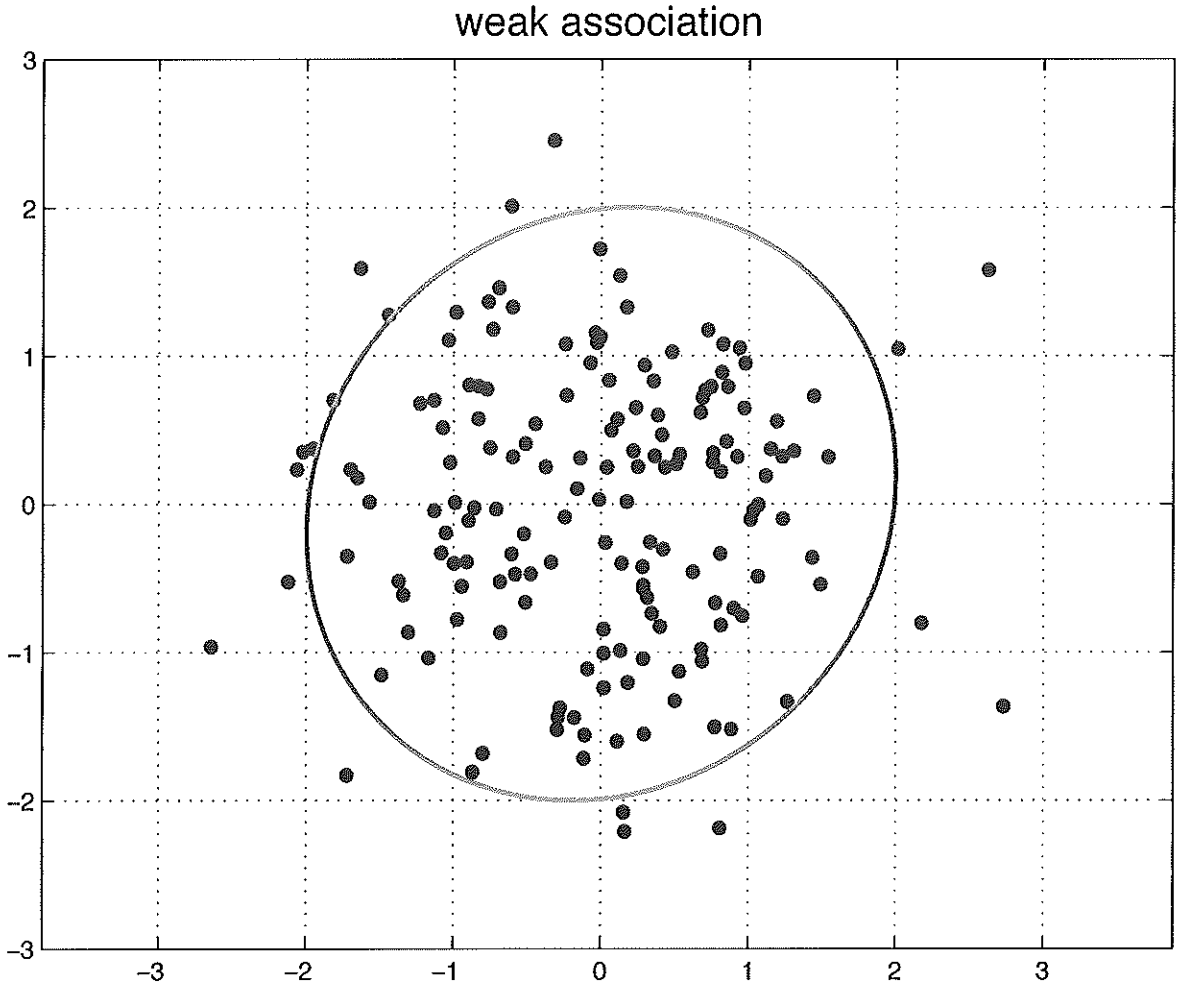
positive correlation

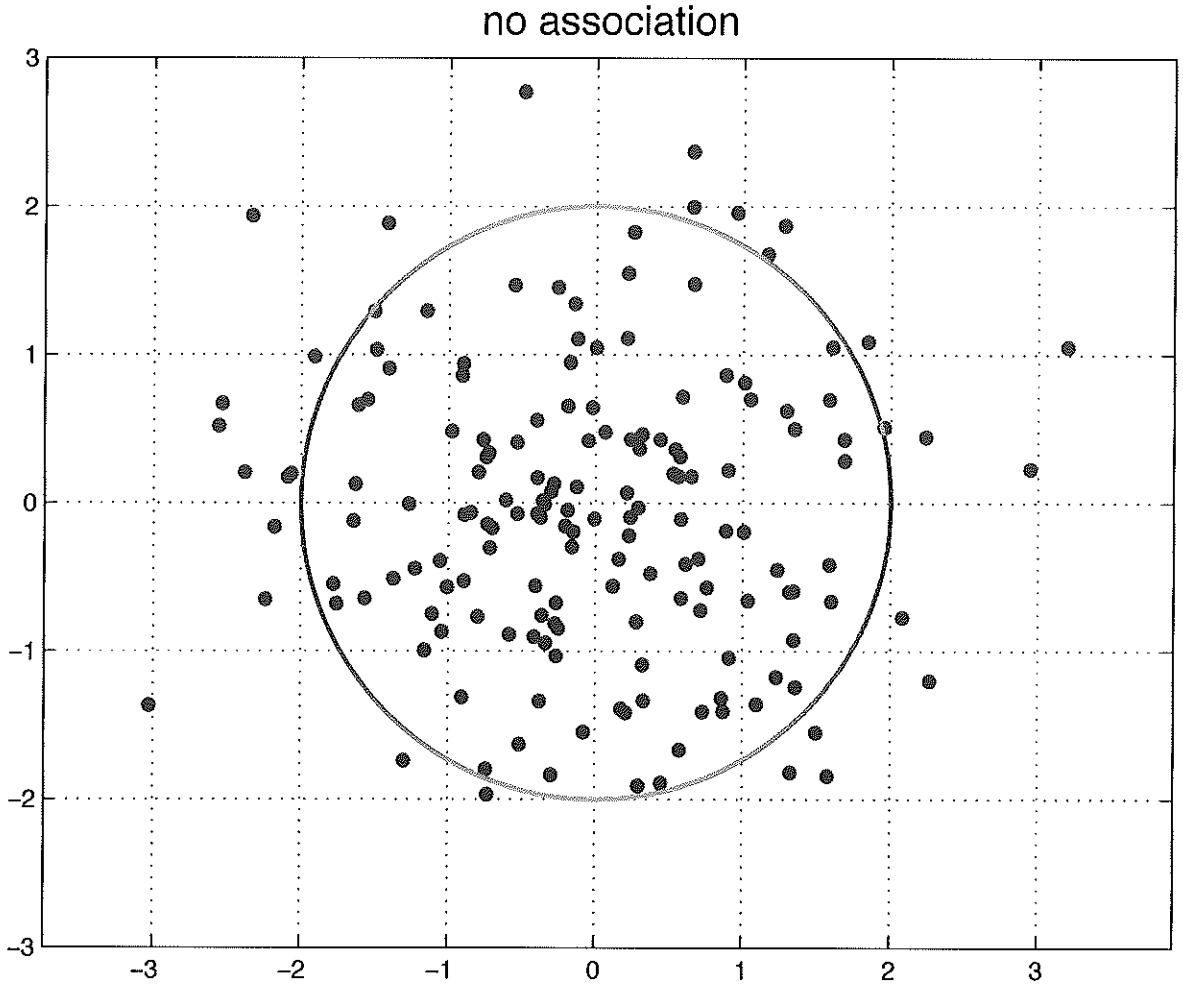
students with higher quiz scores tend ~~to~~ to do better on the midterm.

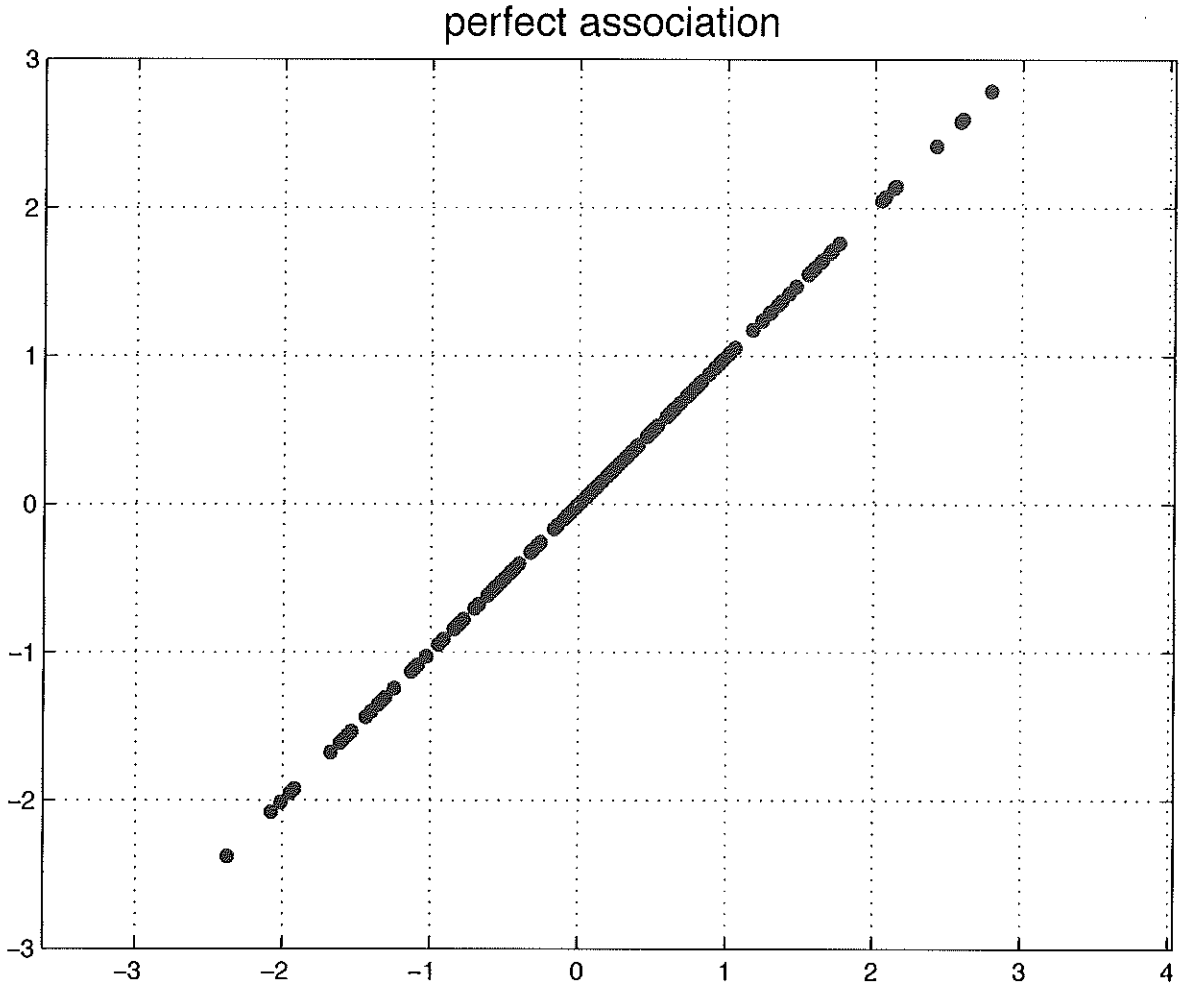
but still a lot of spread.

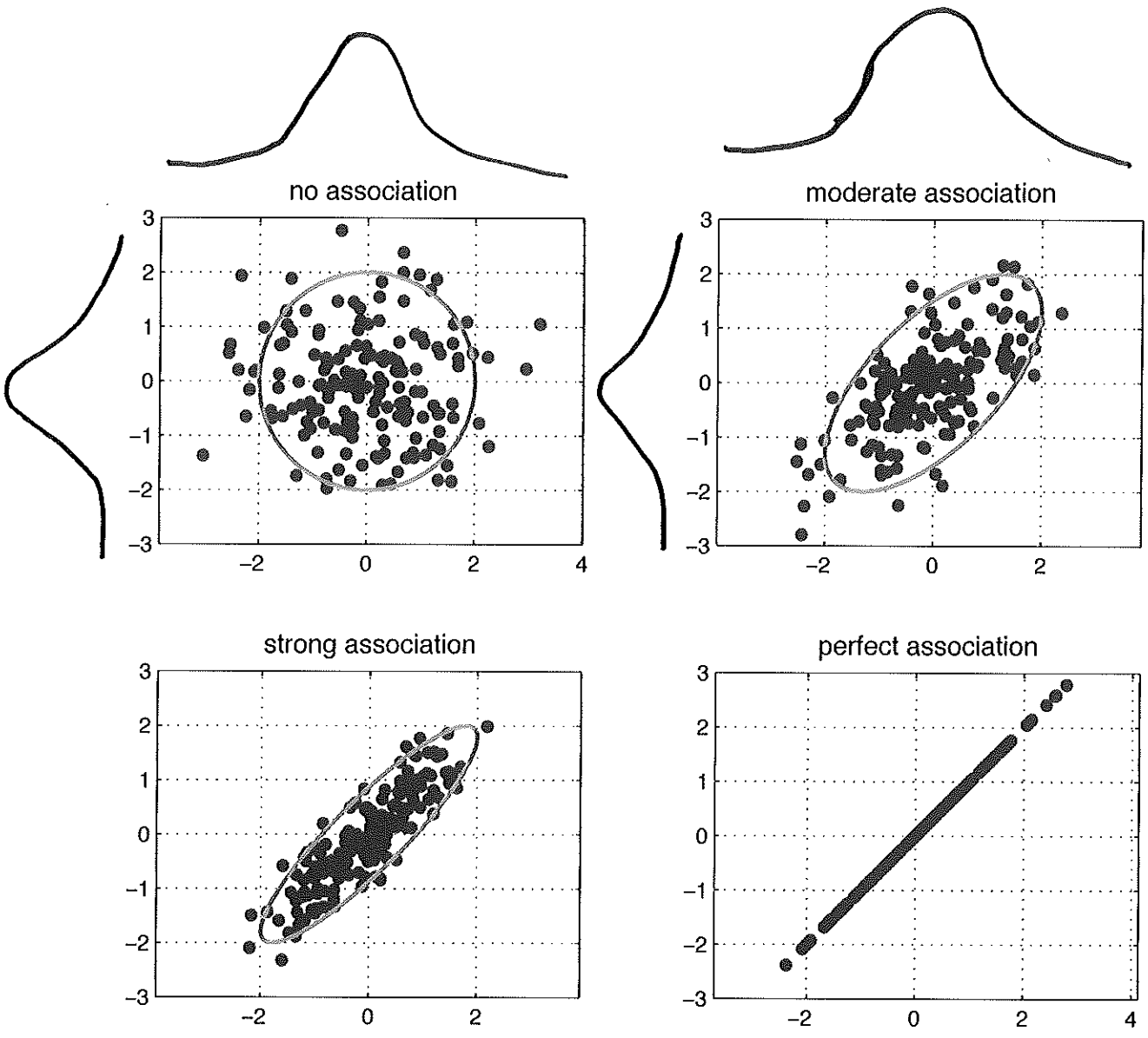

strong association
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


moderate association
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


weak association
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


no association
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


perfect association
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4



Mean$_x$, SD$_x$

Mean$_y$, SD$_y$

are the <u>same</u> for all 4 cases.

If association is strong, knowing the value of one variable helps a lot in predicting the other.

Conversely, if the association is weak, knowing one doesn't help much in predicting the other.

How do we measure the strength of the association

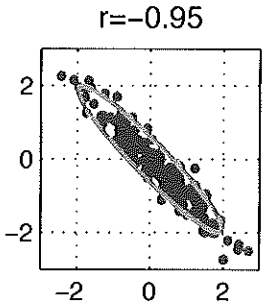
For 1D, we summarised our data by

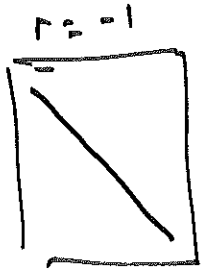
Mean + SD

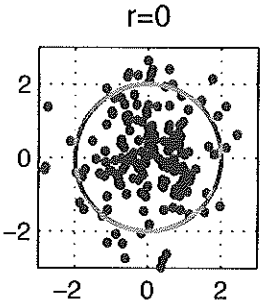
knowing $mean_x$, SD_x, $mean_y$, SD_y is not enough.

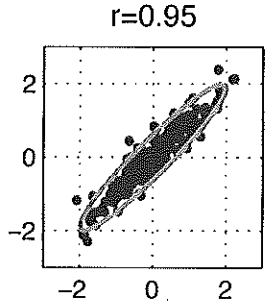
Also need correlation coefficient measures how tightly clustered the points are.

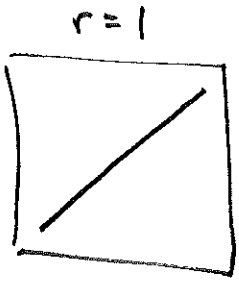
$$-1 \leq r \leq 1$$


r=−0.95
2
0
−2
−2
0
2


r = -1


r=0
2
0
−2
−2
0
2


r=0.95
2
0
−2
−2
0
2


r = 1

Computing the correlation coefficient

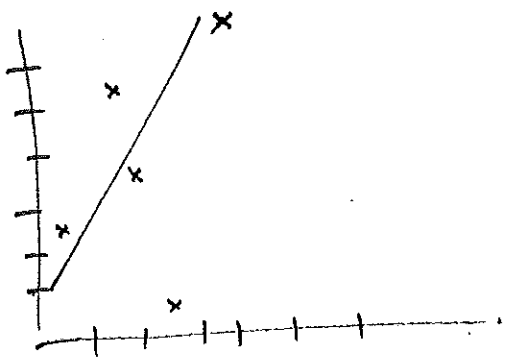
1/ convert each variable to standard units

2/ compute the average of the products

$$r = \text{average of}\left(\begin{array}{c}x \text{ in standard} \\ \text{units}\end{array} \times \begin{array}{c}y \text{ in standard} \\ \text{units}\end{array}\right)$$

eg.

x	y	x in std units	y in std units	product
1	5	-1.5	-0.5	0.75
3	9	-0.5	$+0.5$	-0.25
4	7	0	0	0
5	1	$+0.5$	-1.5	-0.75
7	13	$+1.5$	$+1.5$	2.25



Mean $x = 4$

$$SD_x = \sqrt{\frac{(1-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (7-4)^2}{5}}$$

$$= \sqrt{(9 + 1 + 0 + 1 + 9)/5}$$

$$= 2.$$

$$\text{Mean } y = \frac{35}{5} = 7$$

$$SD_y = 4$$

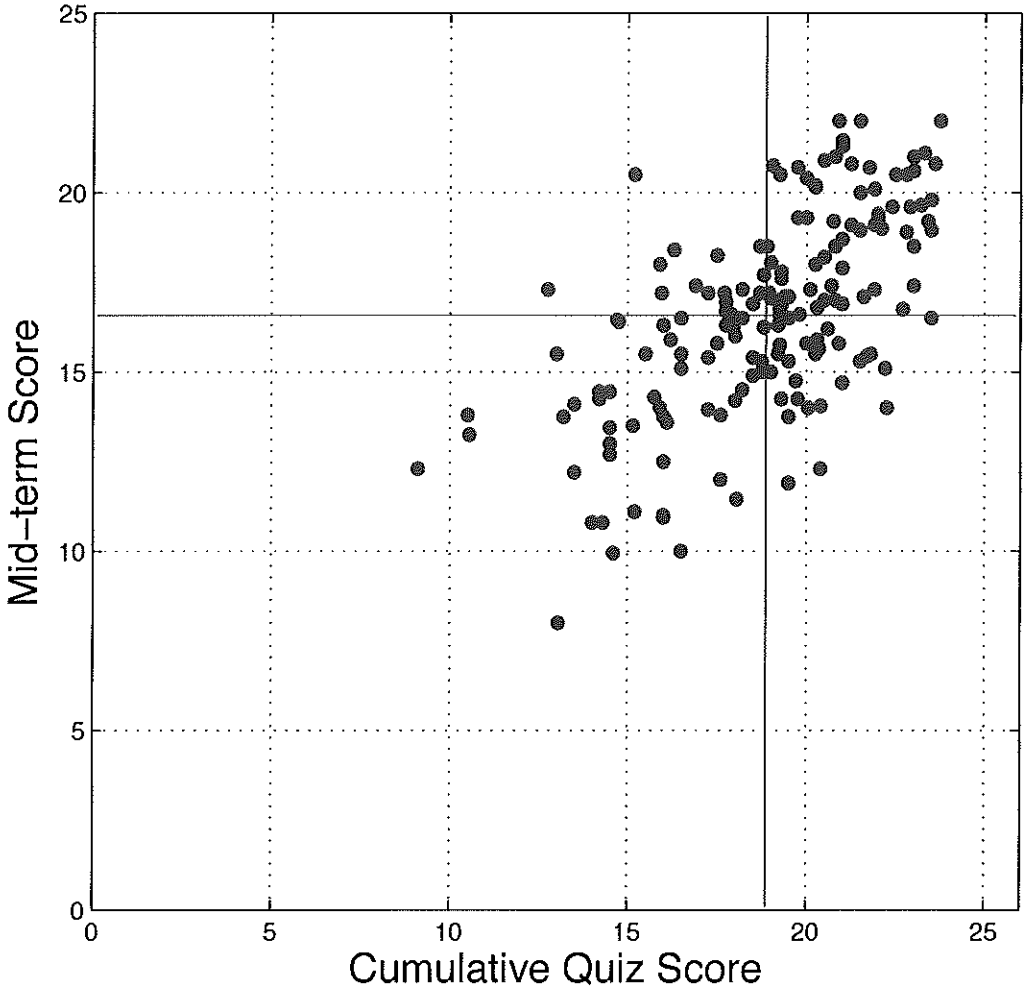
r = average of products of

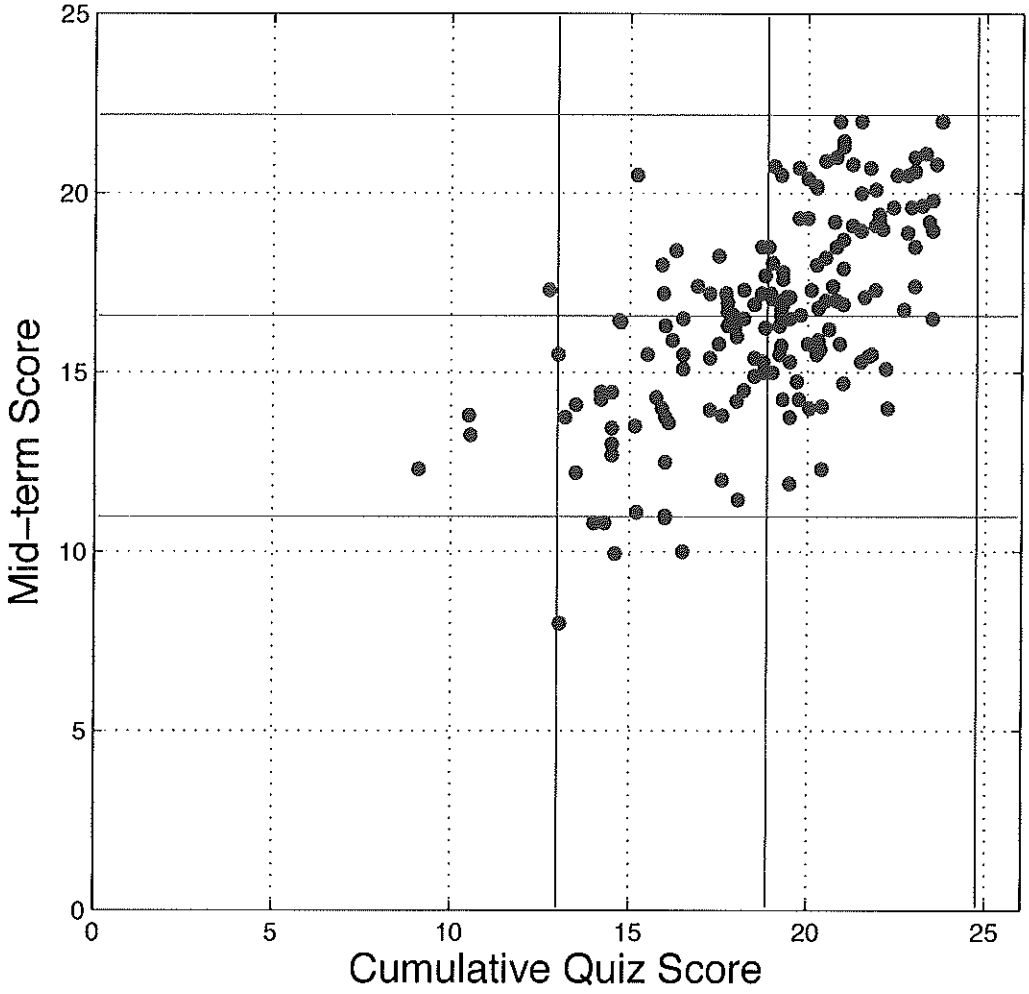
x in std units + y in std units

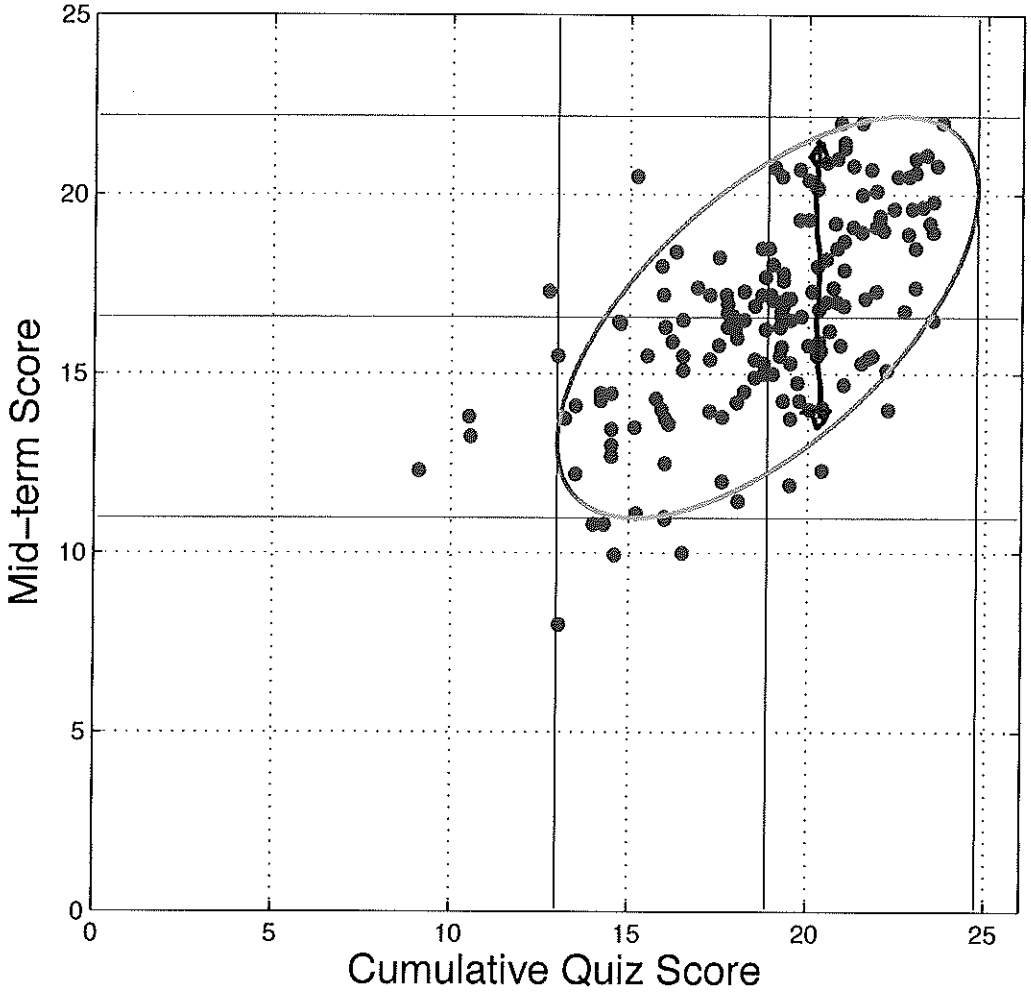
$$r = \frac{0.75 - 0.25 + 0 - 0.75 + 2.25}{5}$$

$$= \frac{2}{5} = \underline{0.4}$$

There is some correlation, but there is also a fair amount of spread in the data.

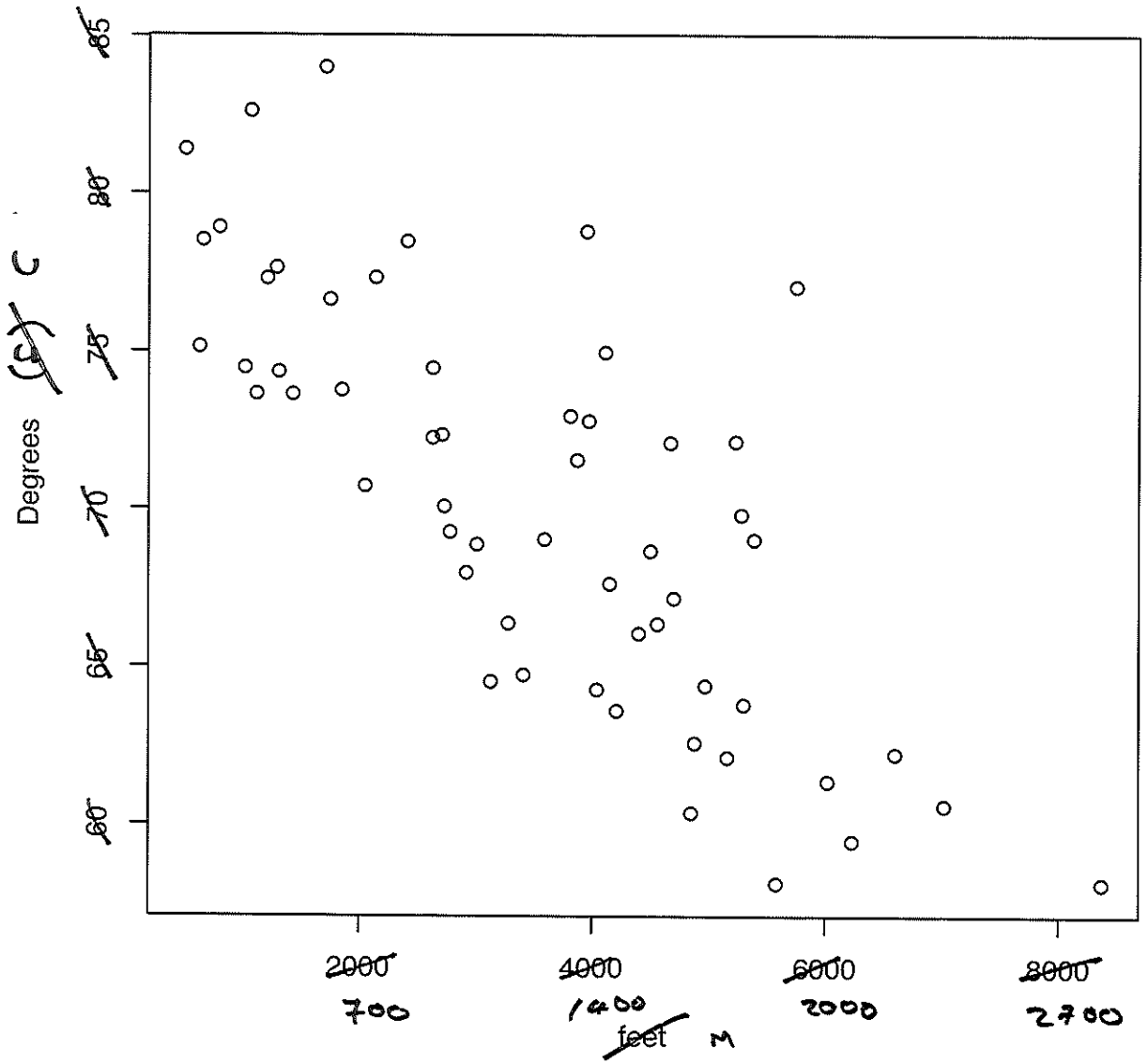

Mid–term Score
Cumulative Quiz Score
0
5
10
15
20
25


Mid–term Score
Cumulative Quiz Score
0
5
10
15
20
25



$r = 0.63$

Moderate degree of correlation.



The correlation coefficient does not depend on the units of measurement.

It does not change if we measure heights in M rather than ft.

Features of the Correlation Coefficient

- it's a pure number — it has no units (inches, pounds, dollars, etc.)

- because we converted x and y to standard units

consequences?

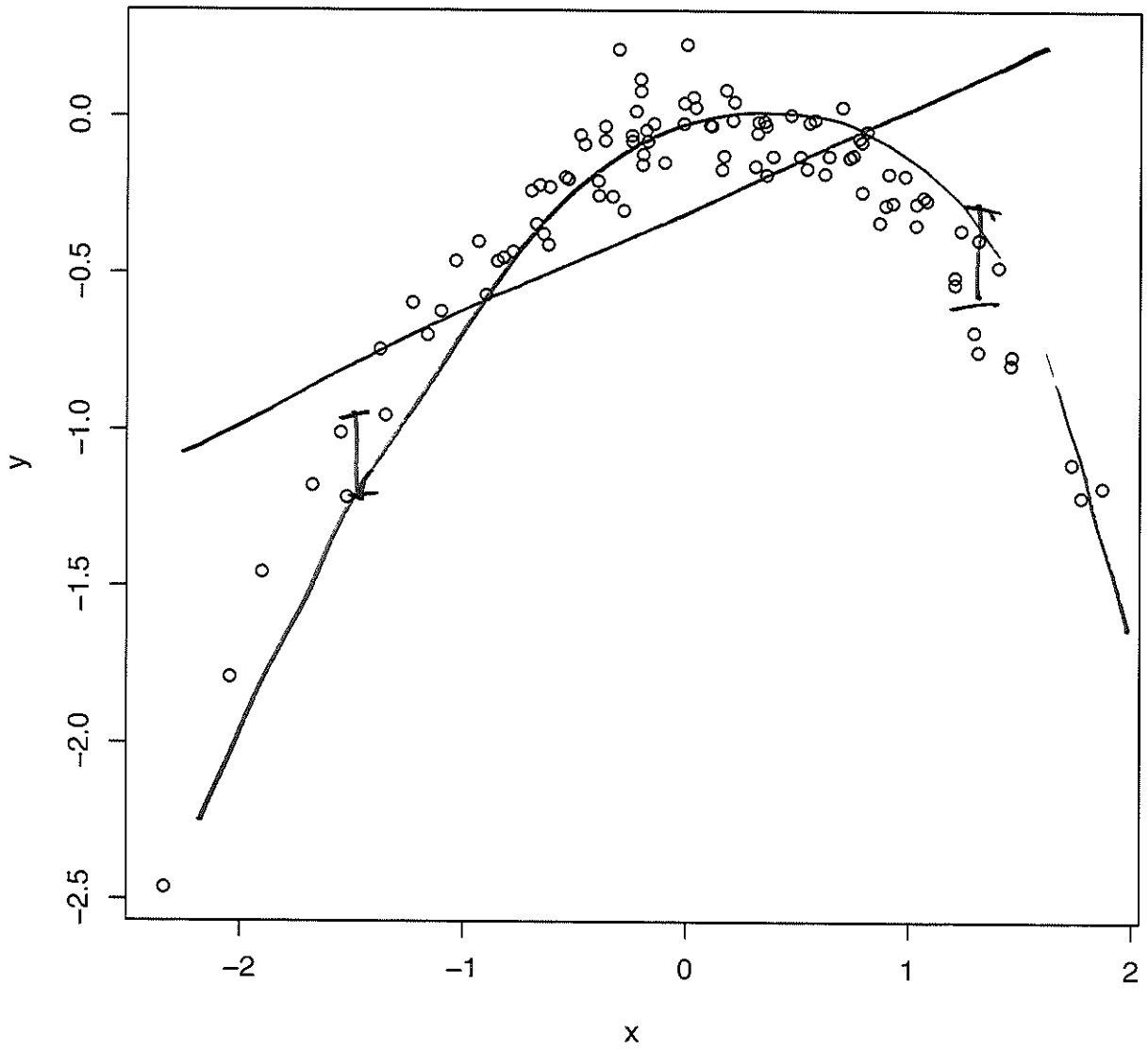
We can change the units of measurement without changing the correlation coefficient.

changing units involves offset + scale factor

When convert to standard units, we remove the effect of these offsets + scales

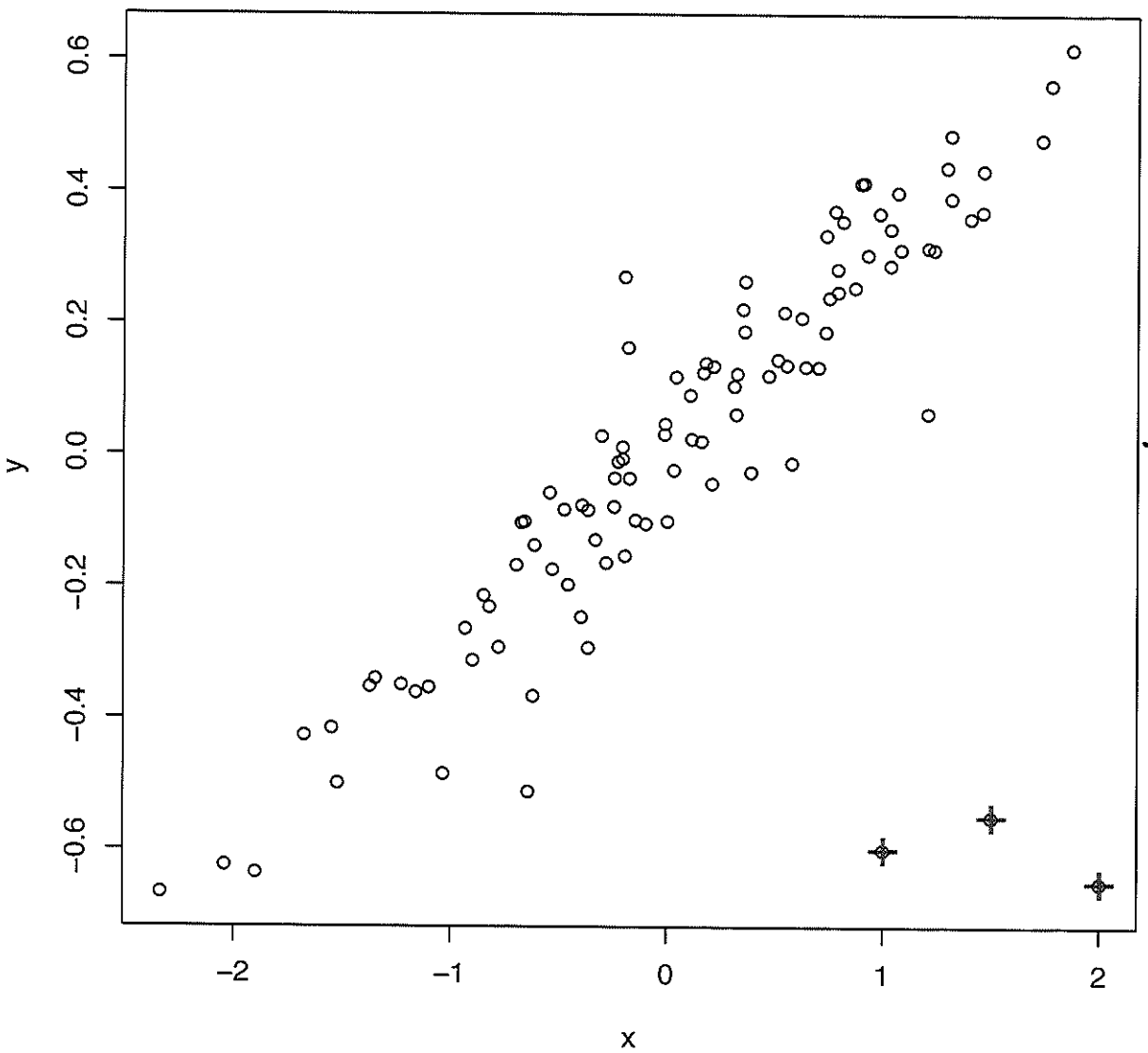
→ values in standard units do not depend on the units of measurement

⇒ correlation coefficient doesn't either.



$r = 0.3$

- very low because the association is non-linear.



correlation coefficient

including the outliers

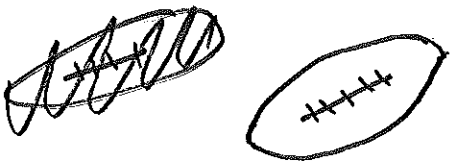
$r = 0.75$

excluding the outliers

$r = 0.944$

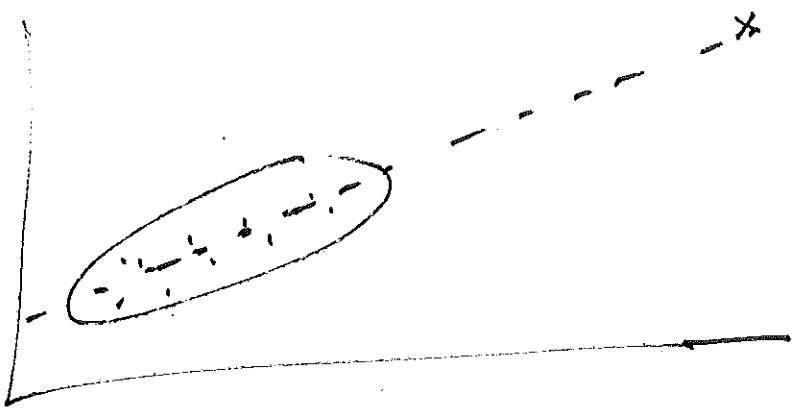
Correlation coefficient measures linear association.

if scatter diagram has the oval shape ("football shaped")



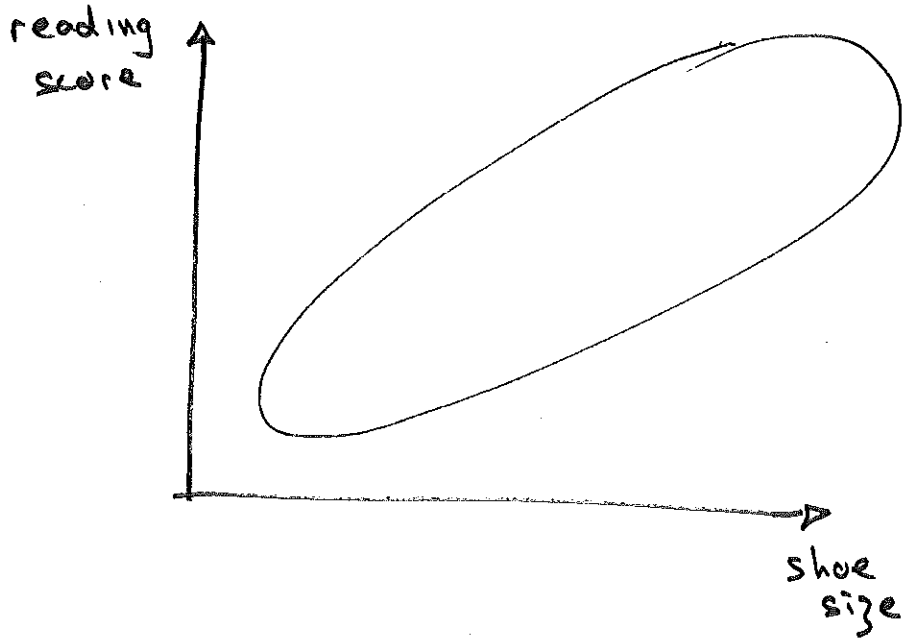
Doesn't apply to non-linear association.

r is sensitive to outliers.



Need additional information to justify removing outlies.

Association is not causation



age as a confounding factor.