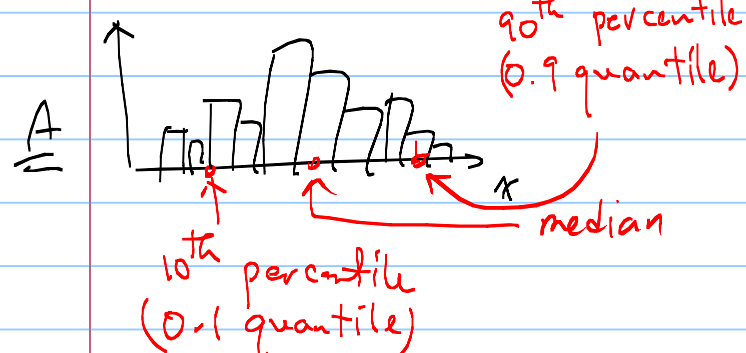
# Lecture 9 (Ch. 2-3)

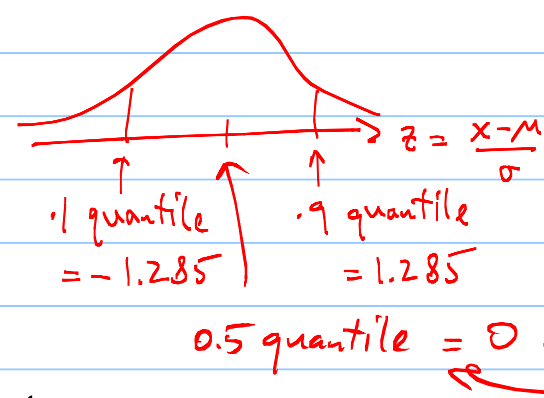
The business of estimating pop. params from sample stats refers to any distr. E.g., one says that $\bar{x}$ and $s$ provide point estimates of $\mu$ and $\sigma$ of the normal dist. IF the data come from a normal dist. to begin with.

Q: But, how do we know if our data come from a Normal dist?

Easier Q: How do we know if our data come from std. Normal?

A: Compare sample quantiles (of data) with distr. (or theoretical) quantiles.




**Example:** (Very Crude!) Here is (sorted) data:

$-1, +1, 3, 4, 4.5, 5, 5.5, 6, 6.5, 8, 9$

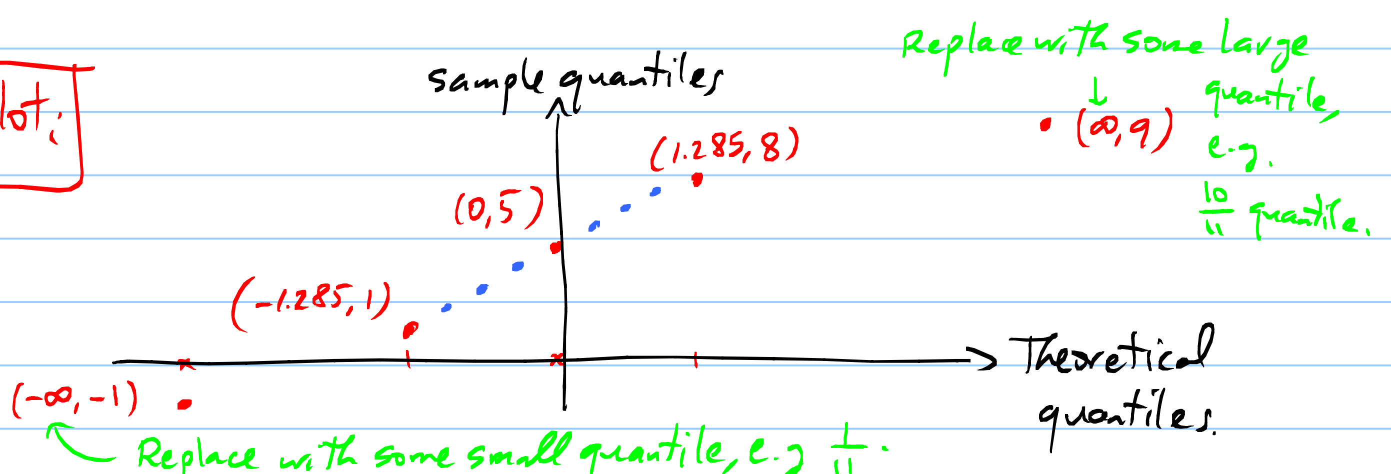
$-1$ = 0th quantile, $+1$ = 0.1 quantile, ... $5$ = 0.5 quantile ... $8$ = 0.9 quantile, $9$ = 1.0 quantile

→ I.e. the 0.1 sample quantile is +1, etc.

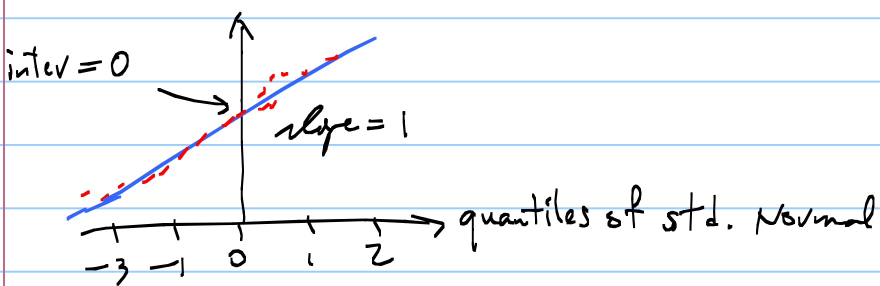
→ Theoretical quantiles: The 0.1 quantile of the std. Normal, etc.

$-\infty$   $-1.285$   ...   $0$   ...   $+1.285$   $\infty$

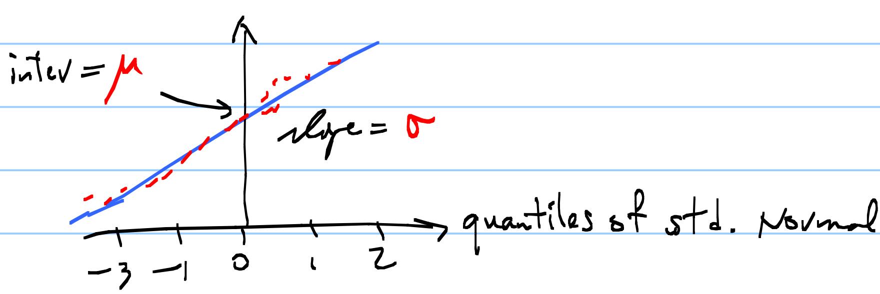
**qq plot:**



If The histogram is consistent with a std. Normal, Then the quantiles/percentiles of data should be equal/comparable to those of The distr.. Then The qq plot should be a straight diagonal line (intercept = 0, slope = 1).



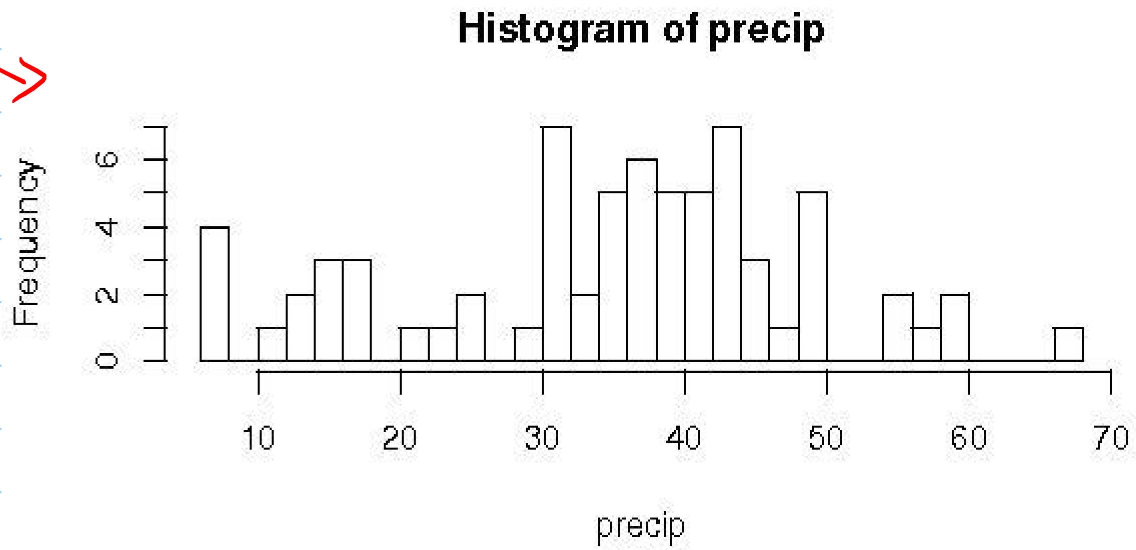

If The data are $\boxed{\text{not}}$ from std. normal, but from $\boxed{N(\mu, \sigma)}$, the only Thing That changes is That The slope becomes $\sigma$, and The intercept becomes $\mu$. NOT too obvious, but pf. in book.




In R: qqnorm(x) where x is The vector of data.

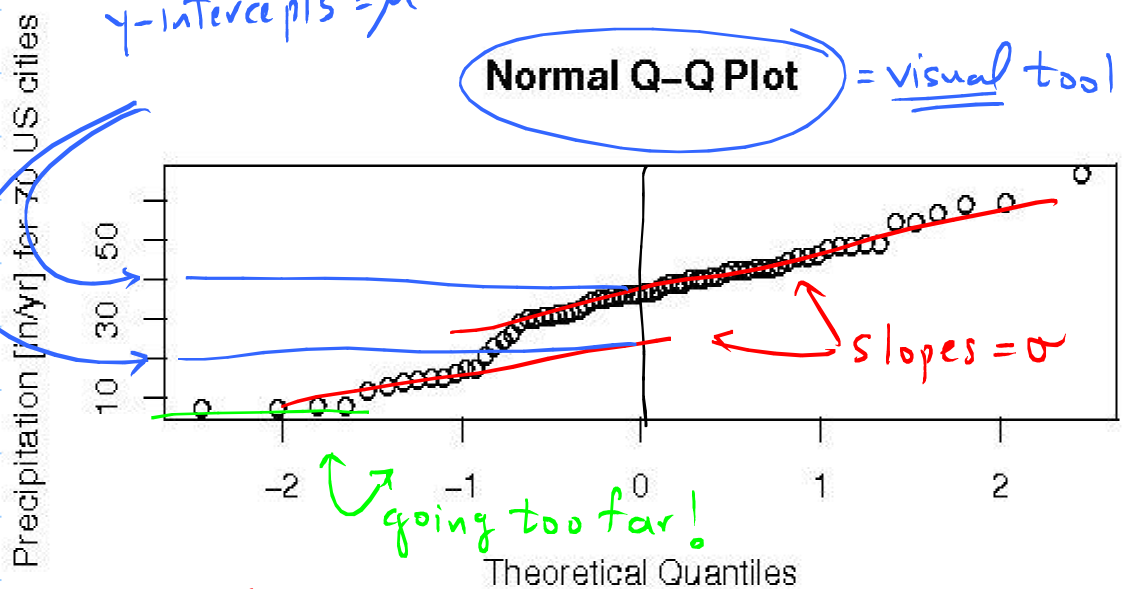
Example

From the histogram, it's hard to tell if the data come from a normal dist., especially because hists depend on bin size.




y-intercepts = $\mu$

= visual tool.




slopes = $\sigma$

going too far!

The plot looks linear, mostly!
So, data are consistent with a Normal.
In fact, it looks like 2 different normals (Bimodal) with diff $\mu$'s, same $\sigma$ (slope).

# Ch. 3

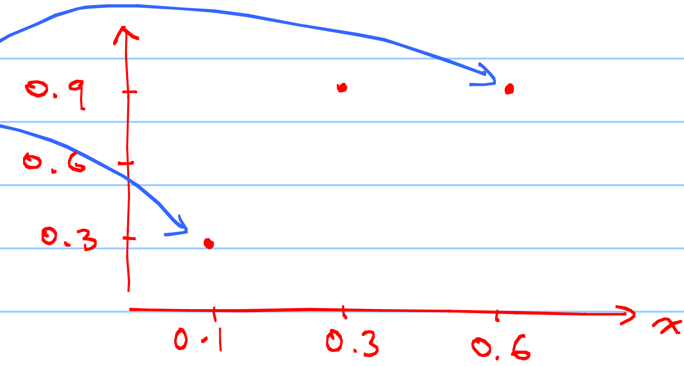
Thus far, our focus has been on 1 column of data, and 1 variable. I.e. univariate analysis.

With 2 (or more) variables, we can do all of the above, but we can also ask about the relationship between them.

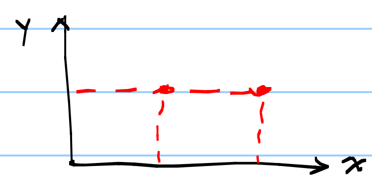
For continuous data: scatterplot                Categ. data, later

Data

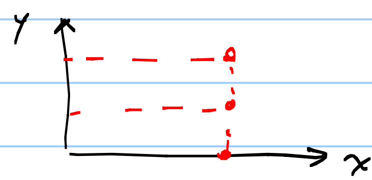
| x | y |
|---|---|
| .1 | .3 |
| .6 | .9 |
| .3 | .9 |
| ⋮ | ⋮ |




Although one purpose of a scatterplot is to summarize and display the relationship between 2 cont. variables, there is nothing that can fully replace it.



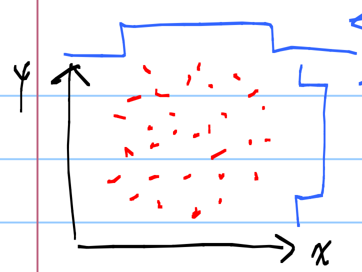

and



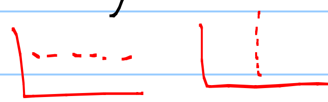

Not unusual. In fact, they are common, (and even necessary)

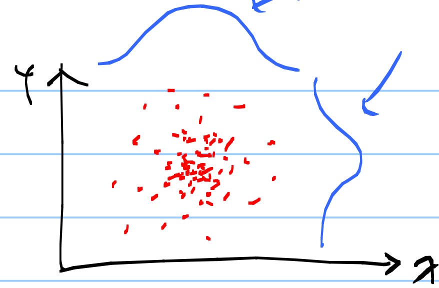
$\boxed{\text{I.e. Given data on 2 vars., do the scatterplot! Of course, histogram each one, too.}}$

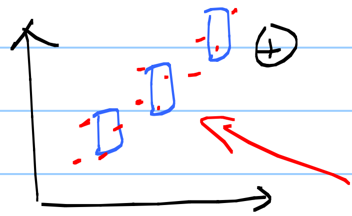
Scatterplot Museum:



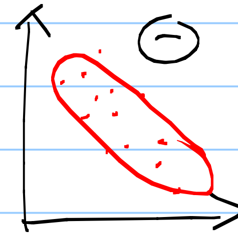

Random association.
(no relation)

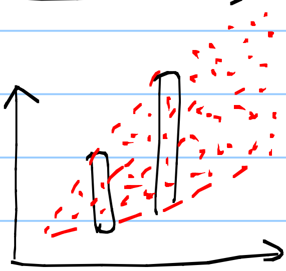




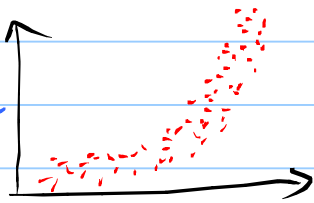




linear, constant variance
y generally increases with $x$,
but y's variance does not.



linear, non-constant variance
var. of y changes with $x$.

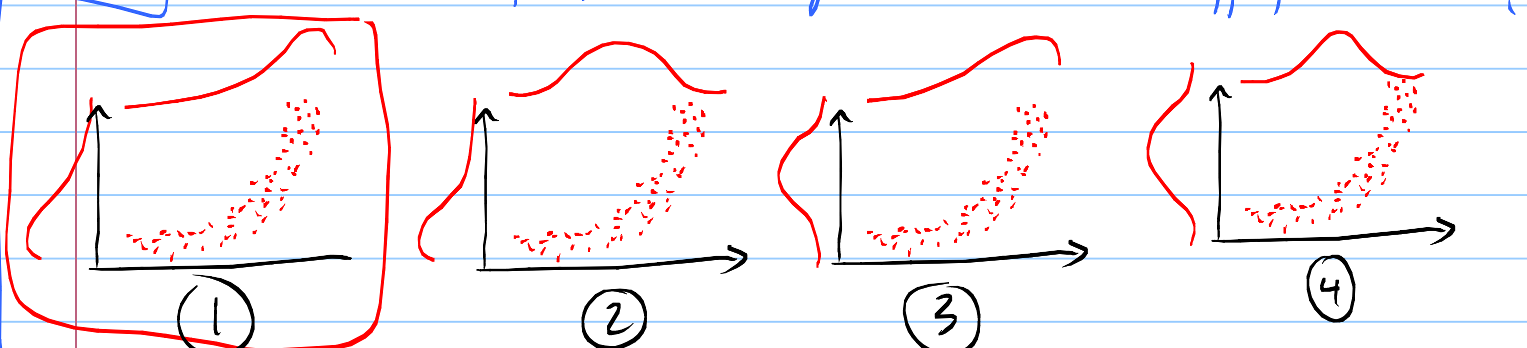


Non linear but monotonic.

Non linear and non-monotonic
(y generally decreases with increasing $x$,
but only up to some point, and
then generally increases with $x$.)

A scatterplot is "the best" device for displaying and studying the relationship (or association) between data on 2 continuous variables.

Q1: For this scatterplot, which fig shows the most appropriate hists?

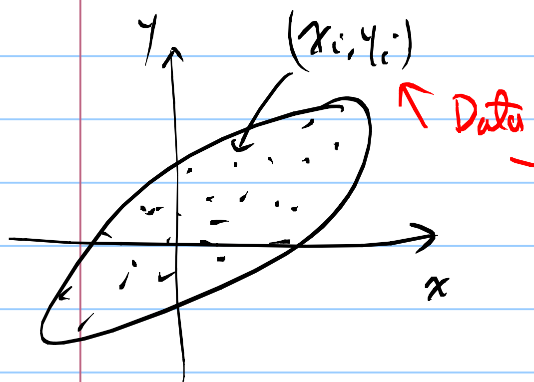


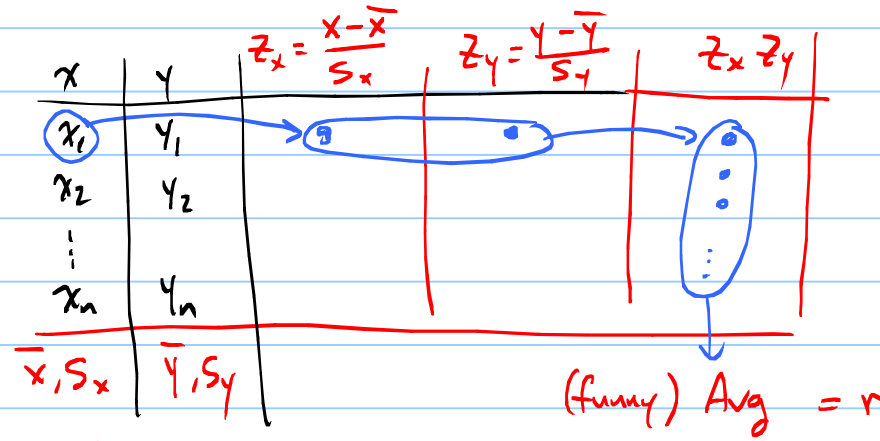
How can we quantify the strength of the associations?

There are many measures of association, the same way there are many measures of "center" or "spread". They capture different facets of "strength."

One popular measure is Pearson's correlation coefficient, denoted $r$ (for sample) and $\rho$ (for population):

like $\bar{x}$ (for sample) $\mu$ (for pop.)




$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z_x} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z_y}$$

$$-1 \leq r < +1$$

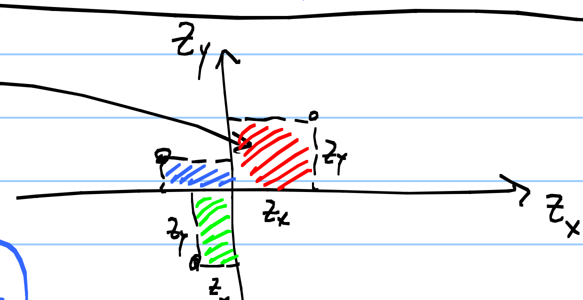
$\boxed{\text{Important:}}$ $r$ measures "skinniness" of scatterplot.

fat scatterplot $\Rightarrow r \sim 0$

skinny " $\Rightarrow r \sim \pm 1$.

But, there are exceptions

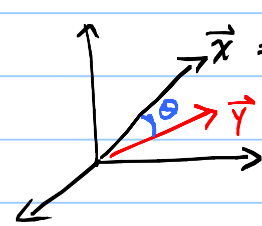
$r =$ Average of "areas"




Only FYI.
Do NOT use on hw/tests

$$r \sim \frac{\vec{x} \cdot \vec{y}}{|\vec{x}|\,|\vec{y}|} \sim \cos(\theta)$$

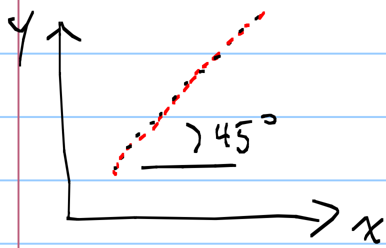
(Recall: $s^2 \sim \vec{x} \cdot \vec{x}$



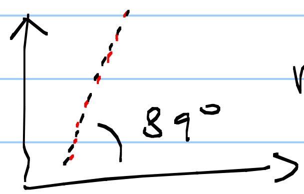

$\vec{x} = \overbrace{(x_1, x_2, x_3, \cdots x_n)}^{n\text{-vector}}$

$\vec{y} = (y_1, y_2, y_3, \cdots, y_n)$

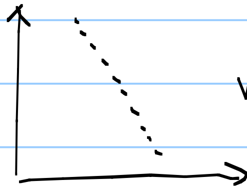
## r museum:



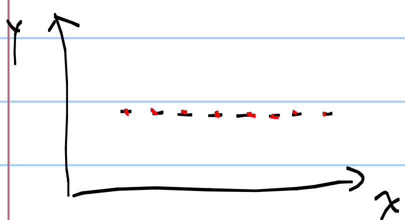

$r = +1$




$r = +1$

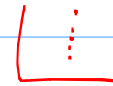


$r = -1$

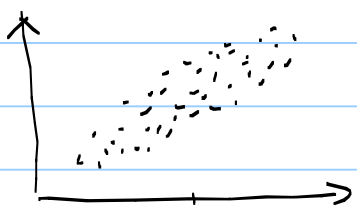



$r = 0$

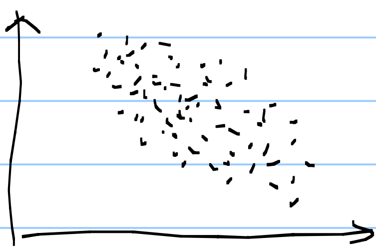
[This involves some limits



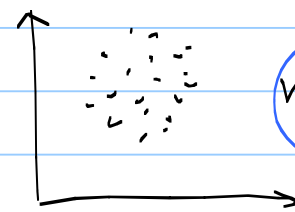
$r = 0$



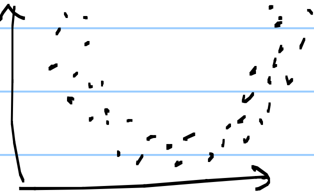
$r \sim 0.7, 0.8$



$r \sim -0.7$
$\sim -0.6$



$r \sim 0$



$r \sim 0$ $\Longrightarrow$ $\boxed{\text{r is a measure of linear association}}$

Important: r is a summary measure of a scatterplot. As such, some info is lost when you look only at r. Look at the scatterplot (too)!

hw-lect 9-1 Do a qq plot of each of the 2 cont. vars. in the data from hw-lect 1. By R. Describe/Interpret the results.

Note: If you find out that there is not much you can say about the qqplot, it may be that your data is not appropriate. This is another chance to correct the error, because later you will be doing more hw problems using your data. So, see me, if you are not sure.

hw-lect 9-2 Make a scatterplot of the 2 continuous vars in hw-lect 1. (By R, or by hand). Describe the relationship.
If it can't be done, see me!

hw-lect 9-3 I gave you a formula that defines r. The book gives two others on p. 110.

a) Start from the formula I "derived" in class, and show that it is equal to

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \qquad \text{I}$$

b) Start from I, and show that it is equal to $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$, where $S_{xx}$, $S_{yy}$, $S_{xy}$ are defined on page 110.

hw-lect 9-4

Suppose n cases of data on x and y fall exactly on the line $y = mx + b$. Compute the value of r.

Hint: In any of the formulas for r, eliminate all y's in favor of x's.

Do not do this

The z's appearing in the formula for r have two nice properties: Their sample mean is zero, and their sample variance is 1. prove these!

I.e. show $\bar{z} = \frac{1}{n}\sum_i z_i = 0$, $\frac{1}{n-1}\sum_i^n (z_i - \bar{z})^2 = 1$