Lecture 12 (ch.3)

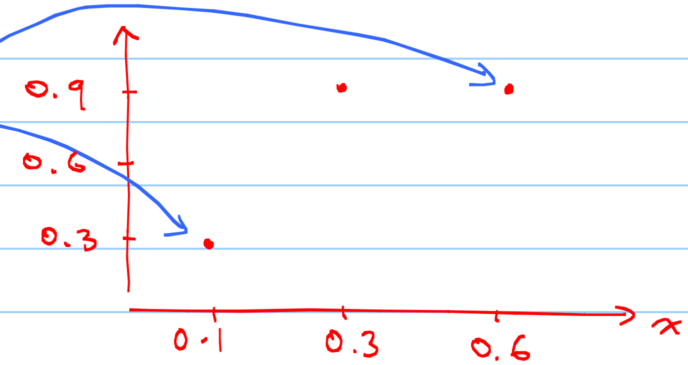
Thus far, our focus has been on 1 column of data, and 1 variable. I.e. univariate analysis.

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

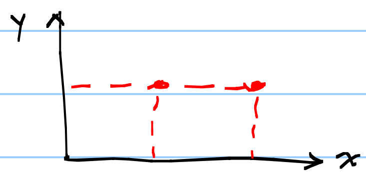
Data

x	y
.1	.3
.6	.9
.3	.9
⋮	⋮

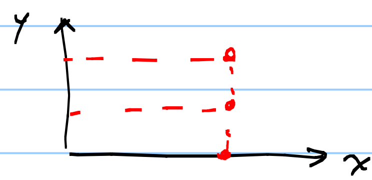


Although one purpose of a scatterplot is to summarize and display the relationship between 2 cont. variables, there is nothing that can fully replace it.

I.e. Given data on 2 vars., do the scatterplot!
Of course, histogram each one, too.

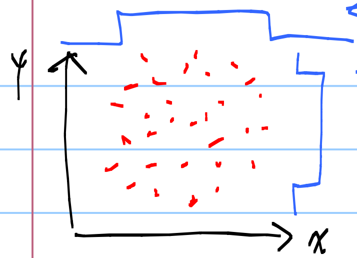


and



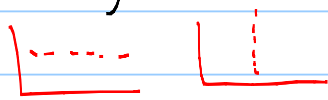
Not unusual. In fact, they are common, (and even necessary)

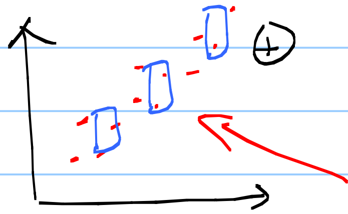
Scatterplot Museum:



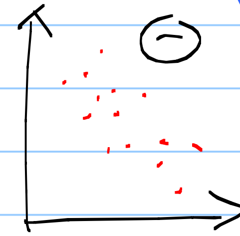
uniform

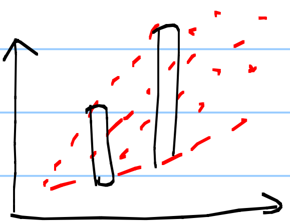
Random association.
(no relation)





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





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

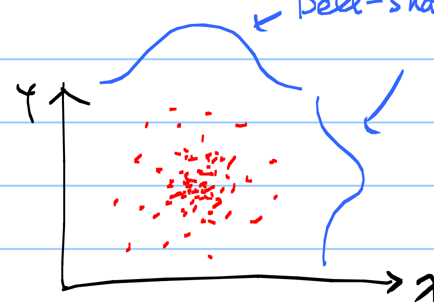
non linear (y decreases with increasing x, but only up to some point. Then reverse.)

generally

periodic x & y.

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

Q1) In the scatterplot shown here,
There is $\boxed{}$ relationship between x & y
A) No, B) Some C) One cannot say.



The diff. with above is the hist of x (and y).

Q Now, how do we quantify the strength of the association between 2 continuous variables?

A There are many measures of strength (like there are different measures of spread of a histogram), and each one captures a different facet of strength.

One popular measure is Pearson's correlation coeff. denoted r (for sample) and ρ (for distribution) ie. population

r gives a point estimate of ρ.

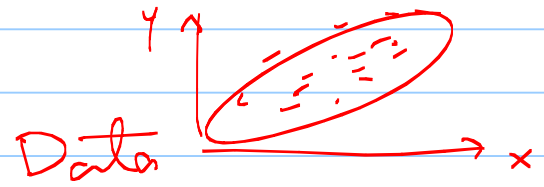
(like $\bar{x}$ gives a point estimate of $\mu_x = E[x]$

sample mean

population mean

Q How do we compute it ?

Data



A

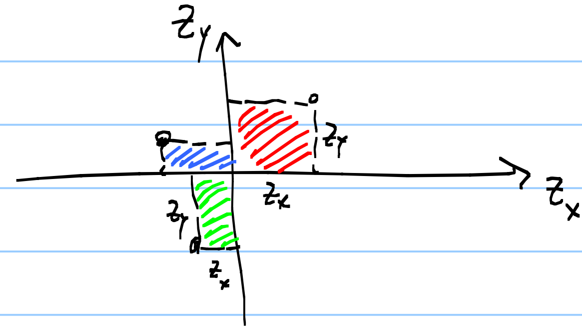
x	y	$z_x = \dfrac{x_i - \bar{x}}{s_x}$	$z_y = \dfrac{y_i - \bar{y}}{s_y}$	$z_x z_y$
x_1	y_1	.	.	.
x_2	y_2	.	.	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n			
$\bar{x}, s_x$	$\bar{y}, s_y$			

$$\frac{1}{n-1} \sum = r . \quad \text{"funny Average"}$$

$$\boxed{r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z_x} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z_y}} \qquad \underline{-1 \leq r_{xy} \leq +1}$$

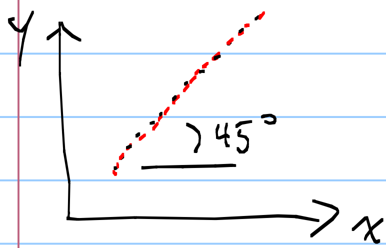
later !

= Average of "areas" $\longrightarrow$

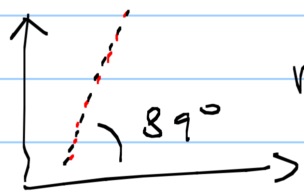


$\boxed{\text{Important}}$: The specific measure of strength that r measures is the "<u>skinniness</u>" of the scatterplot.

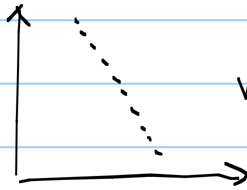
generally $\begin{cases} \text{fat scatterplot} \longrightarrow r \sim 0 \\ \text{skinny} \quad " \quad \longrightarrow r \sim \pm 1 \end{cases}$, But r museum :



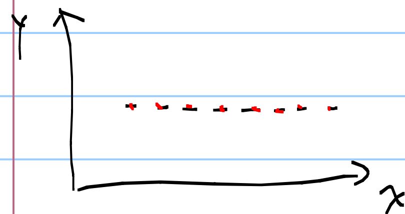
r = +1



r = +1

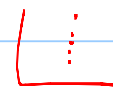


r = −1

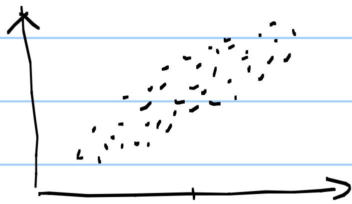


r = 0

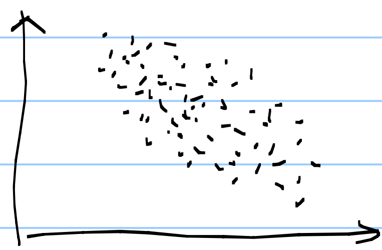
[This involves
some limits



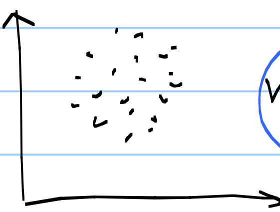
r = 0



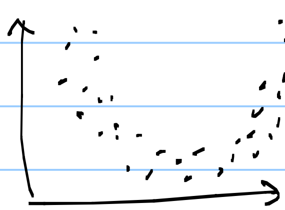
r ~ 0.7, 0.8



r ~ − 0.7
~ − 0.6



r ~ 0



r ~ 0

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_lect12_1 Make a scatterplot of the 2 continuous vars in hw_lect1. (By R, or by hand). Describe the relationship.

If it can't be done, see me!

hw_lect12_2 I gave you a formula that defines r. The book gives two others on p. 108.

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 108.

hw_lect12_3

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 in favor of x.

hw_lect12_4

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