Chp 4 Scatterplots & Correlation

Vocab

response variable: measures outcome (dependent variable); plotted on vertical axis

explanatory variable: may explain (or influence) changes in response variable (independent variable); plotted on horizontal axis

Scatterplot: shows relationship between 2 quantitative variables, where both things are measured on each individual; each individual appears as dot in graph; if have response + explanatory vars., then ex. var on x-axis and res. var. on y-axis; if not, then either variable on either axis.

Note: ☆

- in many studies, goal is to show one variable "causes" another
- relationship between 2 variables can be strongly influenced by other variables that haven't been considered

Ex1 make scatterplot for this data on metabolism.

lean body mass (kg)	36	54	49	42	51	42	40	33	42	35	51
metabolic rate (cal/day)	995	1425	1396	1418	1502	1256	1189	913	1124	1052	1347

M1070

chp 4 (cont)

Association of data

- in a scatter plot, we look at pattern of data, direction of data and strength of the relationship between the variables; outliers will appear to lie outside pattern of data.
- 2 variables are positively associated if best fit line has positive slope (above avg values of one tend to accompany above avg values of other)
- 2 variables are negatively associated if best fit line has negative slope (above avg values of one tend to accompany below avg values of other)
- correlation measures direction and strength of linear relationship between 2 quantitative variables. (usually called r)

collect data (x, y) for each individual

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

r is (avg of products of 2 standardized variables collected for all n individuals)

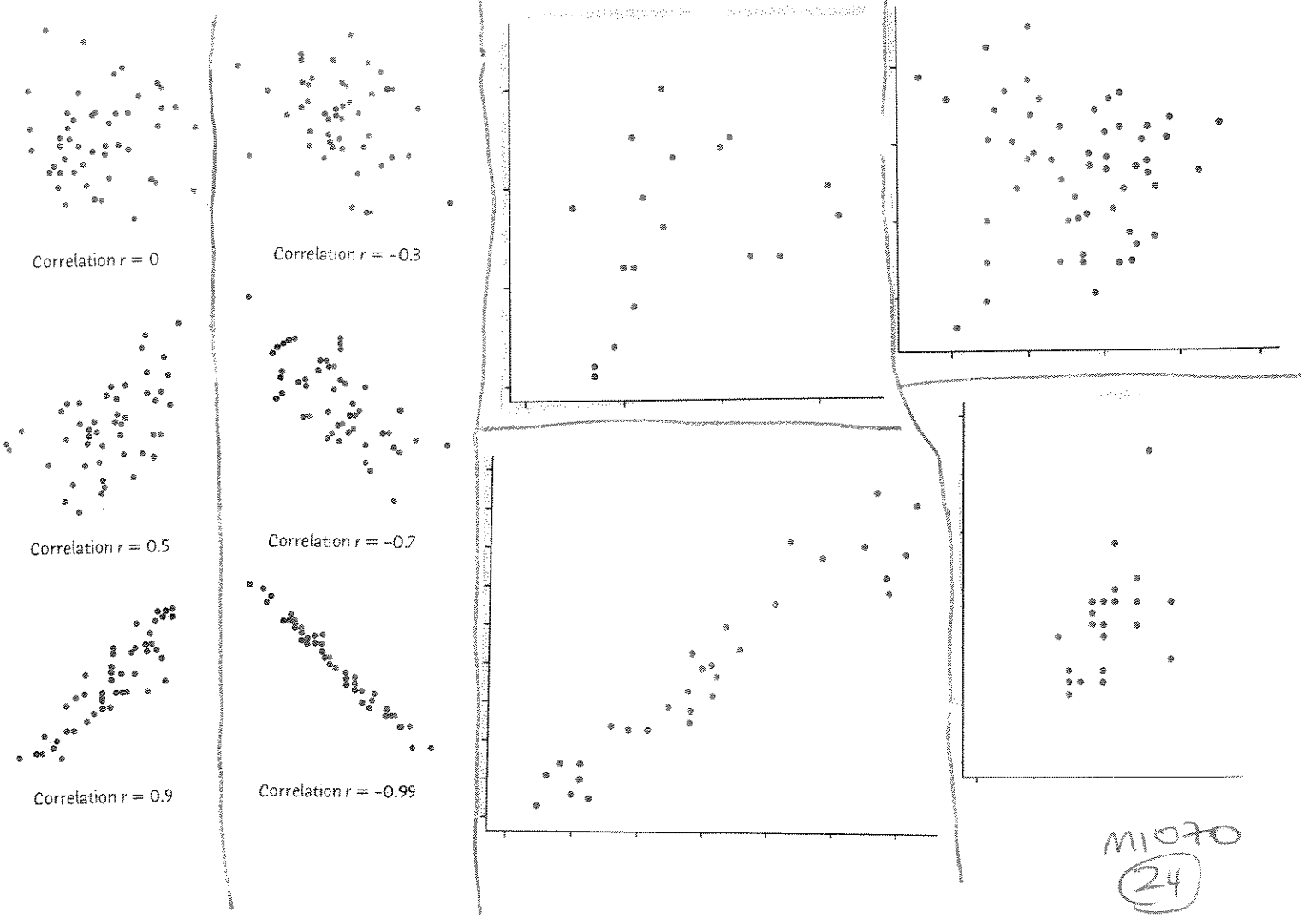
note: ① correlation does not describe curved relationship between variables

② r is not resistant to outliers

Chp 4 (cont)

facts about r

- doesn't matter which variable you assign to be x or y
- r doesn't change if we change units of measurement of x, y or both
- $r > 0 \Rightarrow$ positive association; $r < 0 \Rightarrow$ negative association
- $-1 \le r \le 1$ always
  - $r = 0$ if no linear relationship
  - $r = 1$ if exactly linear relationship w/ positive slope
  - $r = -1$ if exactly linear relationship w/ negative slope



M1070

EX2

Use the scattergrams A-F to answer all of the questions on this page.

Match the following to the scattergram which best fits the description.

_____5. Perfect positive correlation (r=1)

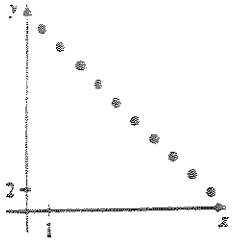
_____6. Strong positive correlation (r = .91)

_____7. Weak positive correlation (r = .3)

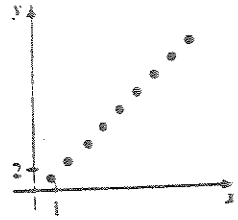
_____8. No correlation (r = 0)

_____9. Strong negative correlation (r = -.85)

_____10. Perfect negative correlation (r = -1)



A



B

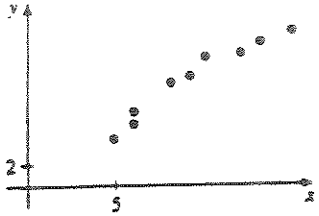
Match the following variables to a suitable scattergram.

_____11. x = infant age in days, y = length in inches

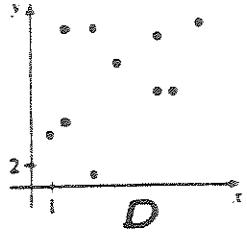
_____12. x = years smoked, y = years you will live

_____13. x = height in cm, y = GPA

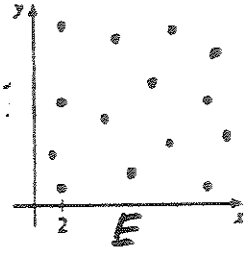
_____14. x = weight in pounds, y = weight in kilograms



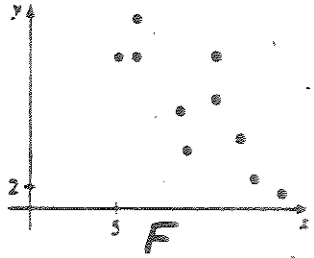
C



D



E



F

Chp 4 (cont)

Ex 3

Ebola and gorillas. The deadly Ebola virus is a threat to both people and gorillas in Central Africa. An outbreak in 2002 and 2003 killed 91 of the 95 gorillas in 7 home ranges in the Congo. To study the spread of the virus, measure "distance" by the number of home ranges separating a group of gorillas from the first group infected. Here are data on distance and number of days until deaths began in each later group:[6]

Distance	1	3	4	4	4	5
Days	4	21	33	41	43	46

(a) Make a scatterplot. Which is the explanatory variable? The plot shows a positive linear pattern.

(b) Find the correlation r step-by-step. First find the mean and standard deviation of each variable. Then find the six standardized values for each variable. Finally use the formula for r. Explain how your value for r matches your graph in (a)

Chp 4 (cont)

Ex 4

Strong association but no correlation. The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	20	30	40	50	60
Mileage	24	28	30	28	24

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.