STATS 1060

Relationships between variables:

Correlation

READINGS: Chapter 7 of your text book (DeVeaux, Vellman and Bock); on-line notes for correlation; on-line practice problems for correlation

NOTICE: You should print a copy of the practice problems and bring them with you to class. Solutions will be reviewed in class and you will have trouble keeping up if you do not have a copy of them with you.

Learning objectives:

1. To properly interpret a scatter plot and correlation coefficient for two quantitative variables.

2. To understand how to think differently about response and explanatory variables.

3. To be able to answer correlation problems 1 to 3 that are provided on-line.

Variables can be associated

RESPONSE VARIABLE: The outcome of a study, or event, you wish to predict

EXPLANATORY VARIABLE: The variable you hypothesize to cause a change in the response variable

- Does high dietary fiber reduce the risk of heart disease?
- Has the incidence of breast cancer been increasing over the last 50 years?
- Does a high fiber diet result in a reduced risk of heart disease?
- Does the concentration of a new drug affect the severity of its side-effects?
- Does alcohol consumption impact ones risk of death due to heart disease?

Relationships between two quantitative variables: Does drinking wine help reduce heart attacks?

Country	Wine consumption[1] (x)	Heart disease[2] (y)
Australia	2.5	211
Austria	3.9	167
Belgium	2.9	131
Canada	2.4	191
Denmark	2.9	220
Finland	0.8	297
France	9.1	71
Iceland	0.8	211
Ireland	0.7	300
Italy	7.9	107
Netherlands	1.8	167
New Zeland	1.9	266
Norway	0.8	227
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
U. K.	1.3	285
U. S.	1.2	199
W. Germany	2.7	172

[1] Wine consumption: liters of alcohol via wine per person per year; [2] Heart disease: Deaths per 10,000 per year

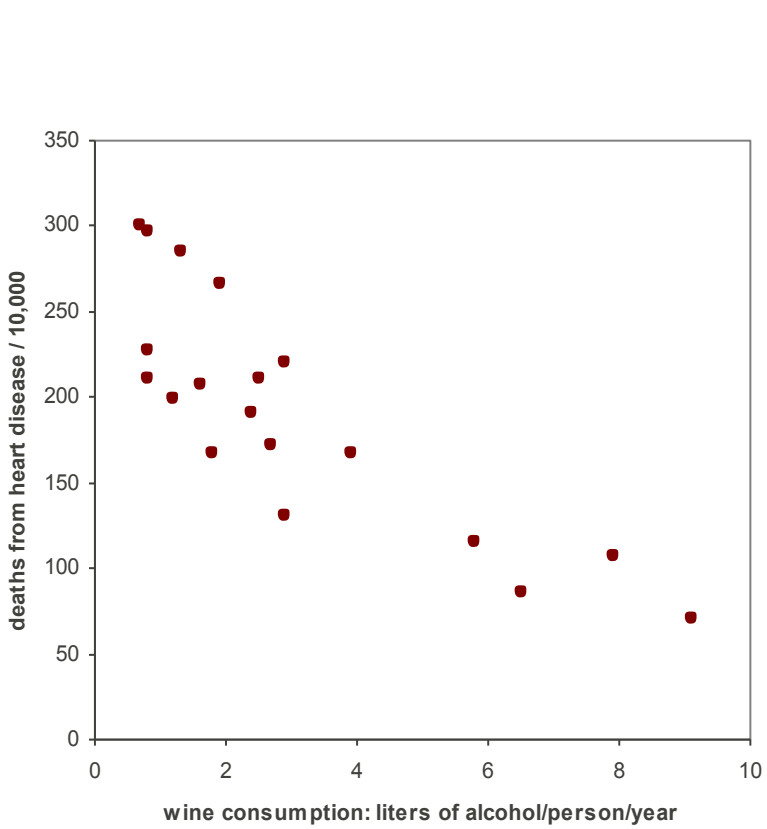
These data will serve as an example for both correlation and regression

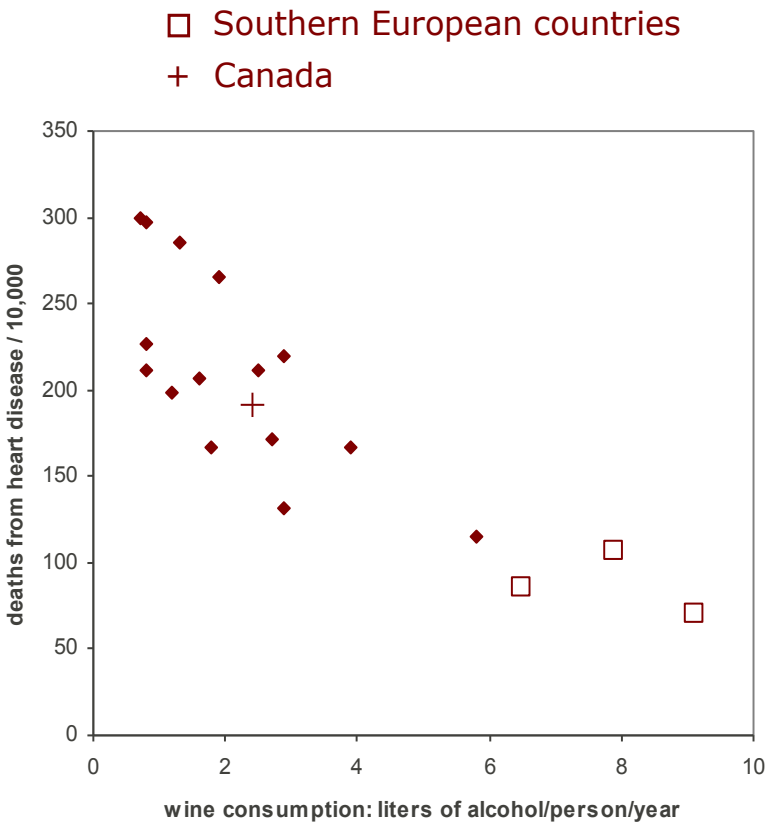
Bivariate observations: Each row corresponds to two variables measured for a single country

Explanatory: Wine consumption measured as liters of alcohol per person per year

Response: Deaths due to heart disease measured as number per 10,000 per year

Scatterplots assist visualization of bivariate data

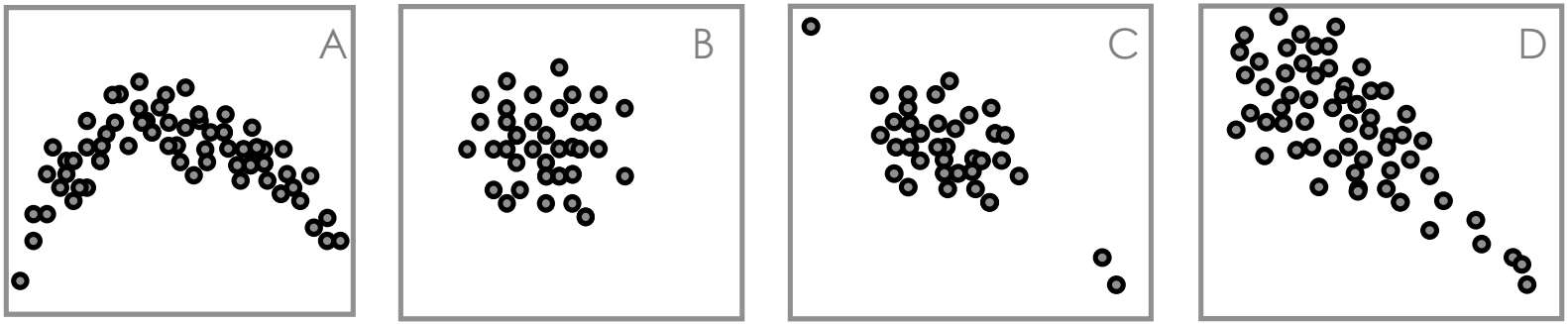




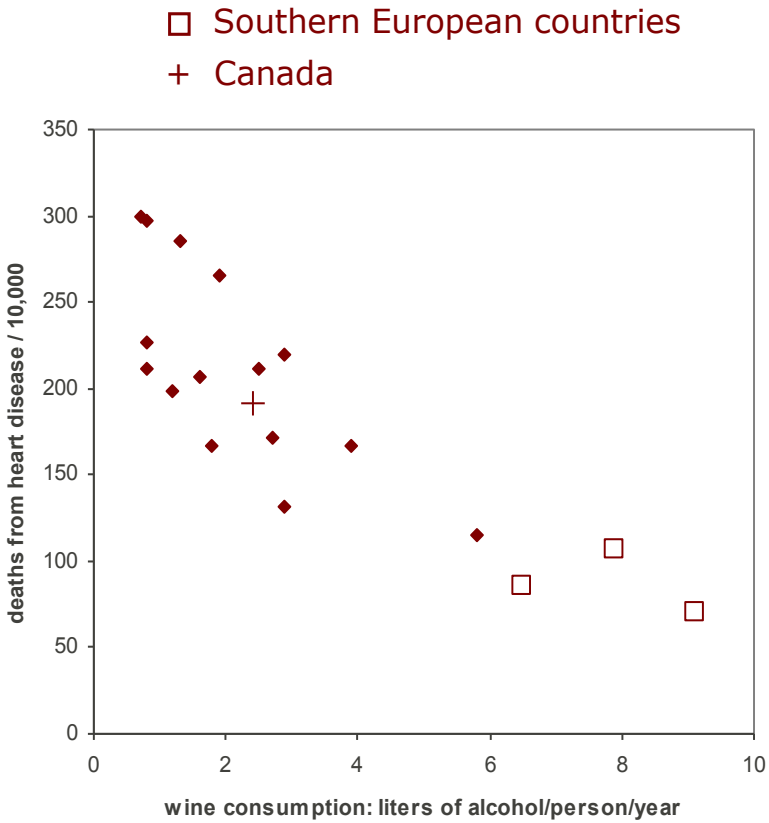
- Each point represents a pair of values (denoted x and y) for a country
- x-axis: explanatory variable
- y-axis: response variable
- Symbols can be used to draw attention to categorical variables

Examine the overall patterns in the data

- form: linear, curved, clustered, something else?
- trend: positive or negative association?
- strength: strong or weak trends?
- outliers: striking deviations from overall pattern?
- variance: does variance in y change with x?

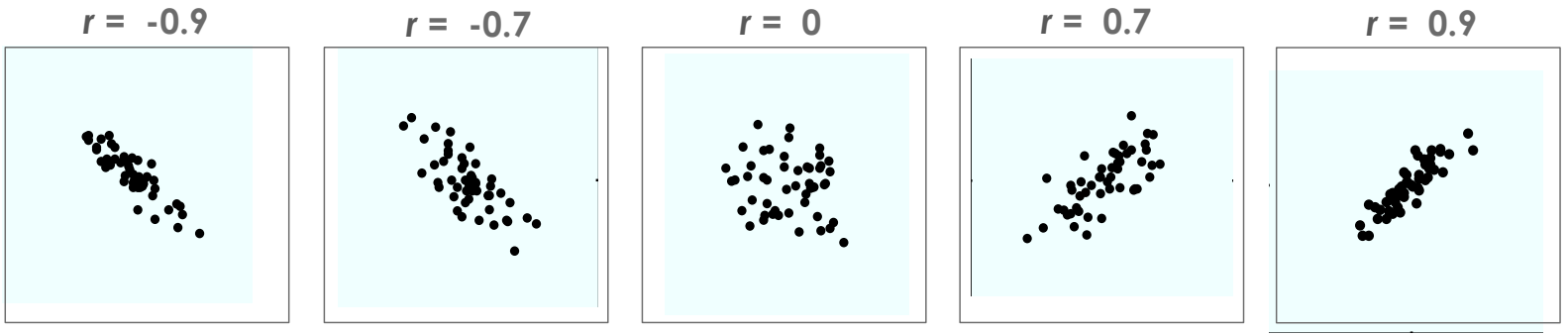


Examine the overall patterns in the data



Is there evidence for the hypothesis that alcohol consumption is associated with the risk of death due to heart disease?

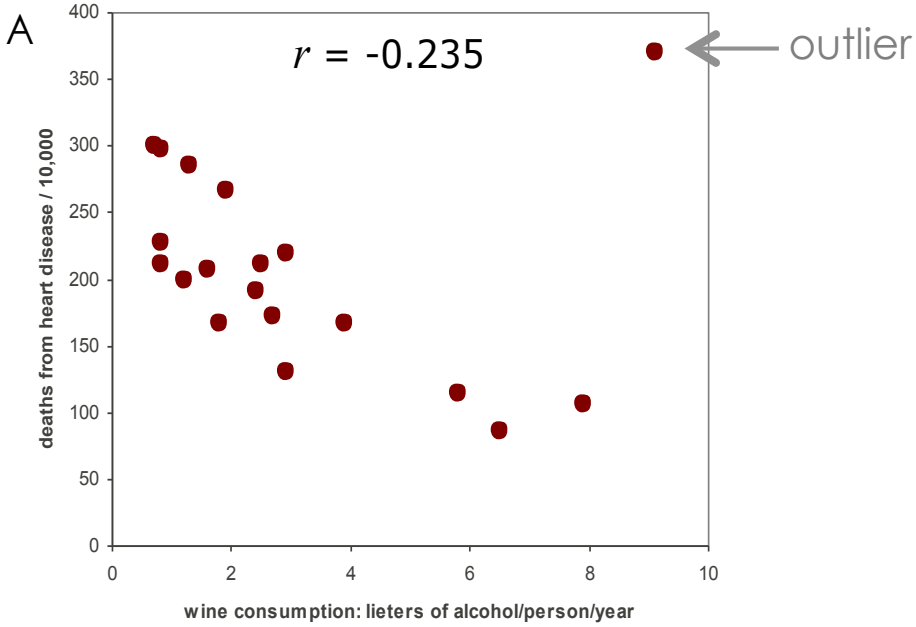
The correlation coefficient (r) is used to measure the strength of a linear association.

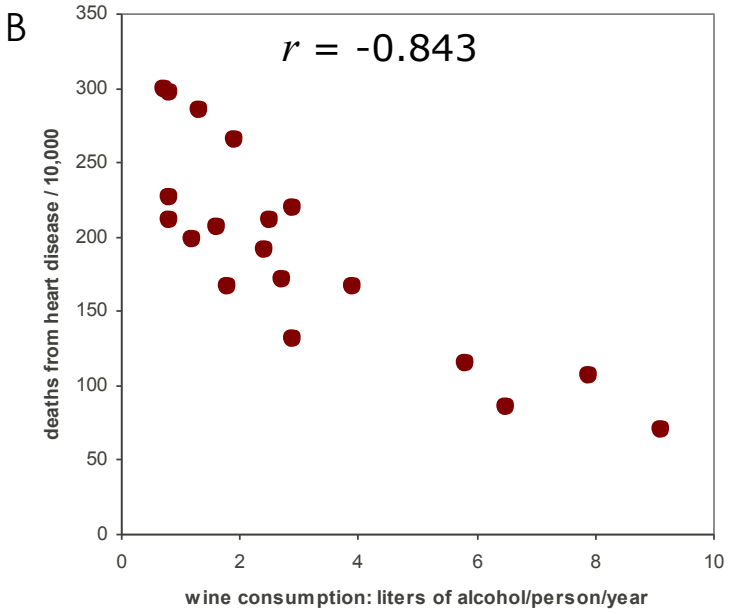


Value of r	Relationship among variables
$r = 1$	The two variables have perfect correlation with no scatter
$r > 0$ (positive)	The two variables tend to increase or decrease together
$r = 0$	The two variables do not vary together in any way
$r < 0$ (negative)	The two variables are inversely related

- x and y must be quantitative variables
- r will always lie between -1 and +1
- r has no units
- r is sensitive to outliers

Caution: The value of *r* can be very sensitive to the presence of outliers





Panel A above shows the scatter plot and correlation coefficient (*r*) when a data entry error produced an outlier (upper right corner of panel A). Panel B shows the change in results obtained when the error is corrected.

Note: outliers need not be due to human error; they can be real values!

The mathematical interpretation of *r*

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

$$r = \frac{\sum z_x z_y}{n-1}$$

- r is based on standardized values (z-scores) for all x and y pairs
- r has no units (recall that z-scores have no units)
- value > mean: + z
- value < mean: - z
- each point in a scatter plot has a (z_x , z_y) pair; their product will be positive or negative
- The strength and sign of r comes from summing over all z_x x z_y

The interpretation of *r*

$$r = \frac{\sum(z_y \times z_x)}{n-1}$$

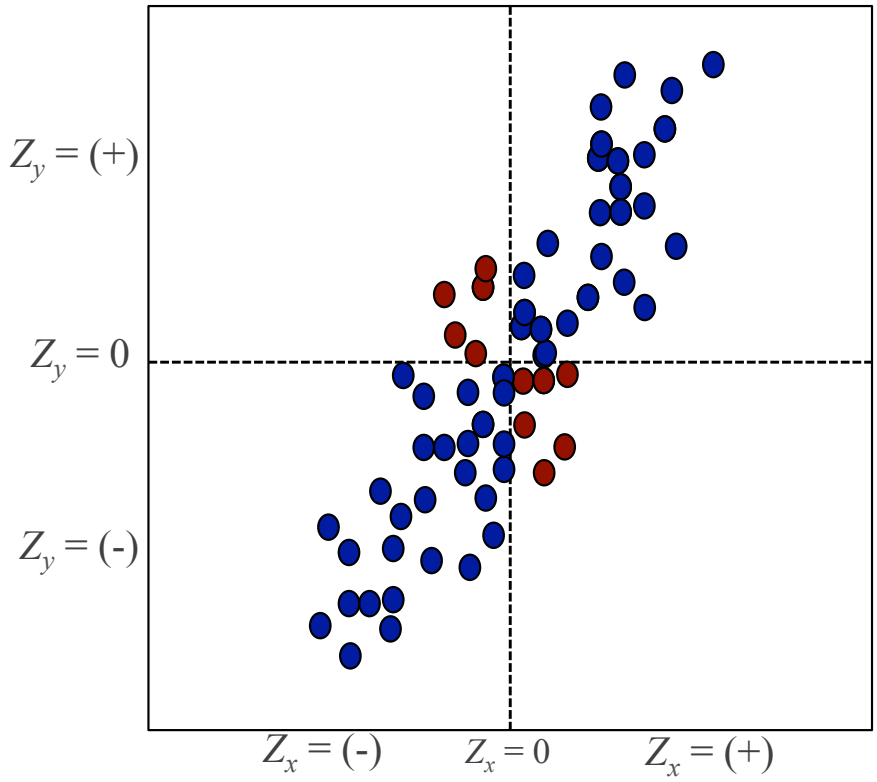
4 quadrants defined by:

$(-z_y \times +z_x) = (-)$

$(-z_y \times -z_x) = (+)$

$(+z_y \times +z_x) = (+)$

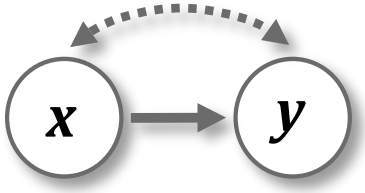
$(+z_y \times -z_x) = (-)$



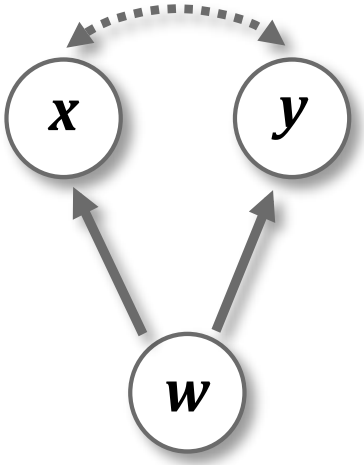
Why is this a positive correlation?

How should we *think* about correlation?

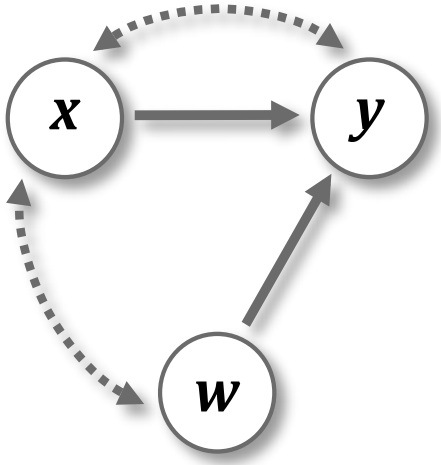
A: causation



B: common response



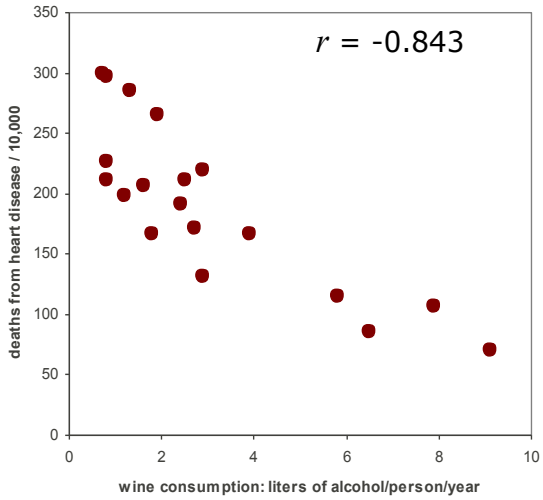
C: confounding

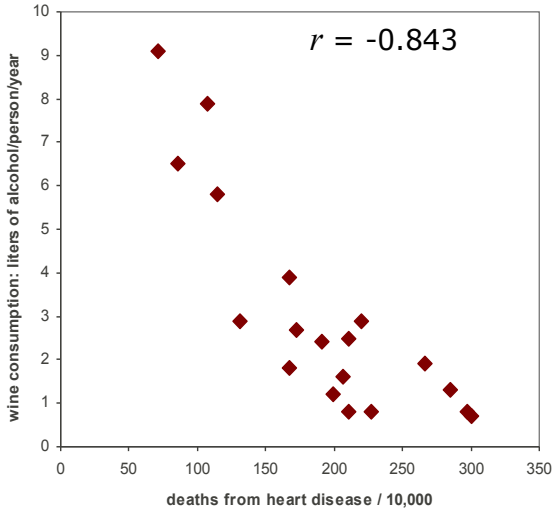


- solid arrow: true cause-effect
- dashed arrow: observed association
- w causes changes to x and y
- w is called a **lurking variable**
- **correlation ≠ causality**

Question: There is a positive correlation between eating ice cream and drowning. Why?

Swapping axes doesn't impact correlation





Does drinking wine cause risk of heart disease to change? Cannot say!

- Drinking wine could reduce risk of death to heart disease
- The rate of heart disease in a country could impact wine consumption
- Both wine consumption and death rate could be impacted by a lurking variable
- The variables could be unrelated, and the result was due to chance

Question: Can you think of any lurking or confounding variables?

Practice problems

The **in-class practice problems** are distributed on-line via the course web site (through Dal's Online Web Learning, or OWL, resource).

Additional problems, and real-time solutions, are provided on line in the form of screencasts. The additional problems are also provided in PDF form via a link on that site. You are strongly encouraged to try working those problems before watching the screencasts. The additional problems will NOT be covered during class time.

Primary URL:
http://awarnach.mathstat.dal.ca/~joeb/Stats1060_Webcasts/Part_1.html

Alternate URL:
http://web.me.com/cadair_idris/stats1060/Part_1.html