PSQF 4143: Section 13

Brandon LeBeau

Correlation Introduction

- Up to now, we have primarily focused on statistics for single variables
- We will now shift to discuss how to quantify and explore the relationship between two variables.
- To display the relationship between two variables, could use a bivariate frequency distribution or a scatterplot.

Bivariate Frequency Distribution

TABLE 5.1-1 Scatterplot of Midparent Height and Height of Adult Offspring[a] (Female Heights Multiplied by 1.08)

	Midparent Height (inches)[b]										
Height of Adult Offspring	*64*	*64.5*	*65.5*	*66.5*	*67.5*	*68.5*	*69.5*	*70.5*	*71.5*	*72.5*	*73*
73.7							5	3	2	4	
73.2						3	4	3	2	2	(3)
72.2			1		4	4	11	4	9	(7)	1
71.2			2		11	18	20	7	4	2	
70.2			5	4	19	21	25	14	(10)	1	
69.2	1	2	7	13	38	48	(33)	(18)	5	2	
68.2	1		7	14	28	(34)	20	12	3	1	
67.2	2	5	(11)	(17)	(38)	31	27	3	4		
66.2	2	(5)	11	17	36	25	17	1	3		
65.2	(1)	1	7	2	15	16	4	1	1		
64.2	4	4	5	5	14	11	16				
63.2	2	4	9	3	5	7	1				
62.2		1		3	3						
61.7	1	1	1			1		1			

[a] Galton (1889, p. 208). I am grateful to Edward W. Minium for bringing these data to my attention.

[b] A circle marks the class interval containing the median of each column.

Figure 1:

Scatterplot

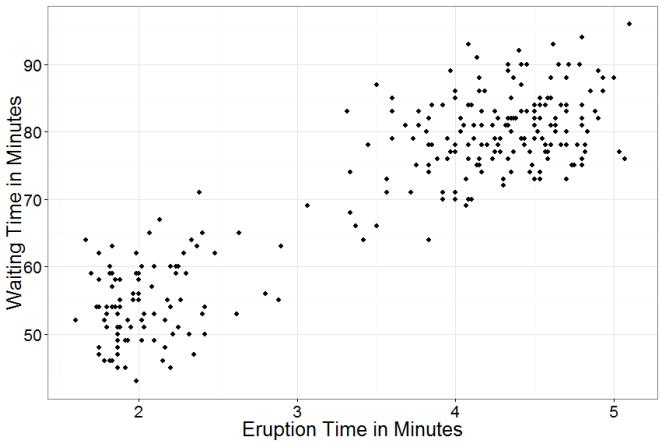


Figure 2: plot of chunk eruptions

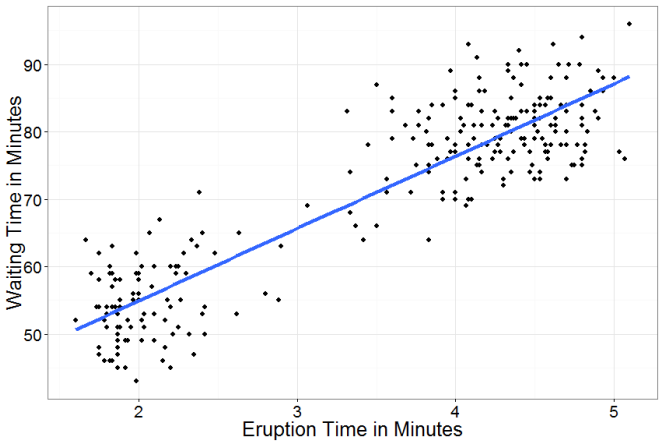


Figure 3: plot of chunk eruptionslinear

Correlation 2

- The correlation gives us a measure to describe the relationship between two variables.
  - The form of the relationship
  - The direction of the relationship
  - The strength of the relationship
- The pearson product-moment correlation coefficient (correlation) describes the amount of **linear** relationship between two variables.
- Notation:
  - Sample: r
  - Population: ρ

Correlation 3

- Correlations can be either positive or negative:
  - Positive: Effort and Achievement – People who expend more effort tend to achieve more.
  - Negative: Cholesterol level and life expectancy – People with lower cholesterol levels tend to live longer.
- A correlation of 1 represents a perfect positive relationship.
  - Example: Annual precipitation in inches and annual precipitation in centimeters.
- A correlation of -1 represents a perfect negative relationship.
  - Example: Number of days in class and number of days absent.
- A correlation of 0 represents no relationship.
  - Example: height and last digit of social security number.

Examples

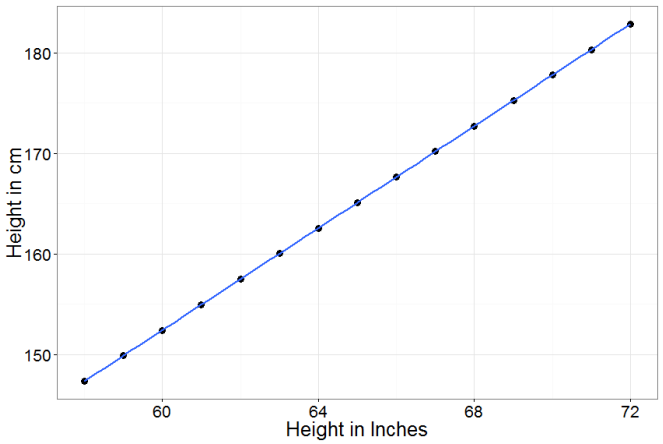


Figure 4: plot of chunk height

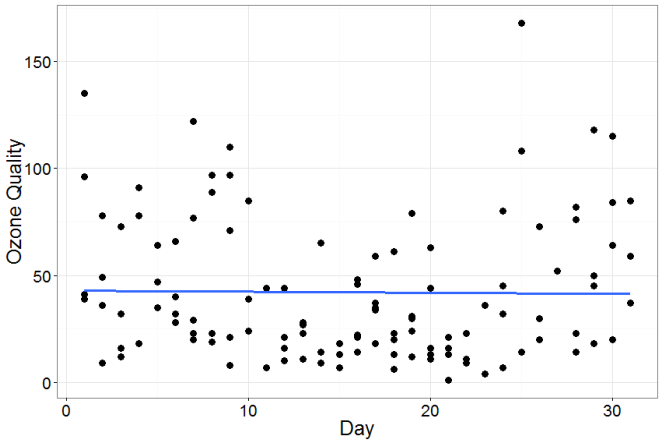


Figure 5: plot of chunk ozone

Real world Examples

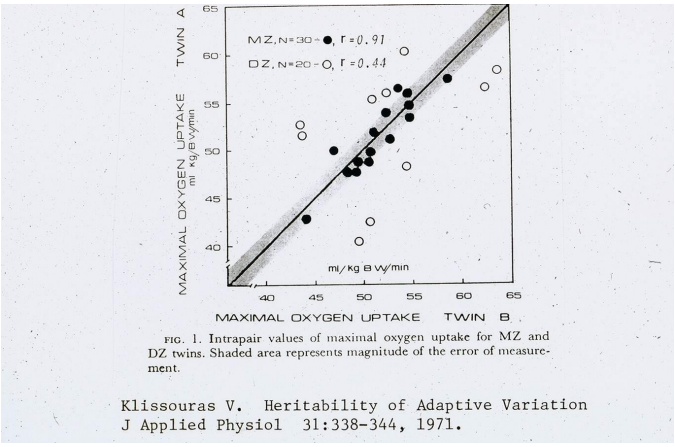


FIG. 1. Intrapair values of maximal oxygen uptake for MZ and DZ twins. Shaded area represents magnitude of the error of measurement.

Klissouras V. Heritability of Adaptive Variation
J Applied Physiol 31:338-344, 1971.

Figure 6: Twins

Real World Example 2

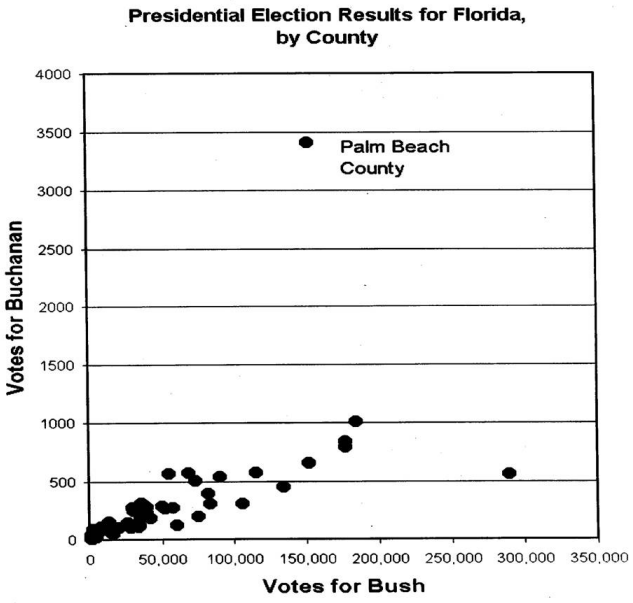


Figure 7: Florida Election

Real World Example 2 cont.

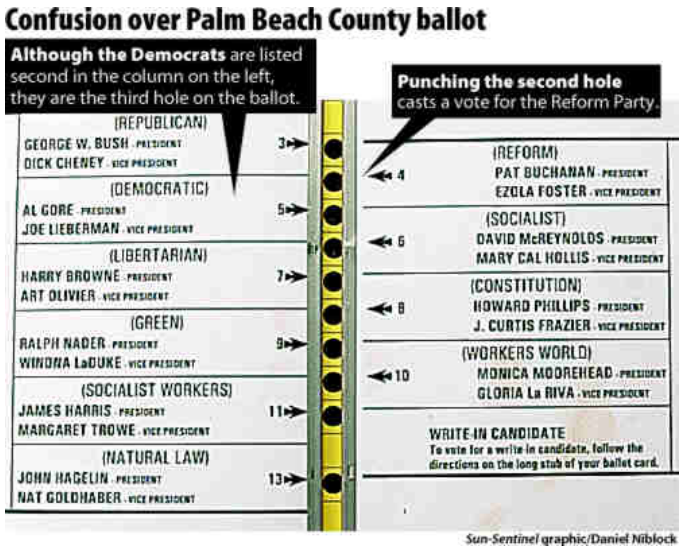


Figure 8: ballot

Guessing Correlations

http://istics.net/stat/Correlations/

Correlation Formula

$$r = \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}}$$

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{s_X s_Y} * \frac{1}{n}$$

where $s_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n}$ is the **covariance**
and s_X and s_Y are the standard deviations of the X and Y respectively.

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

Calculating the Correlation

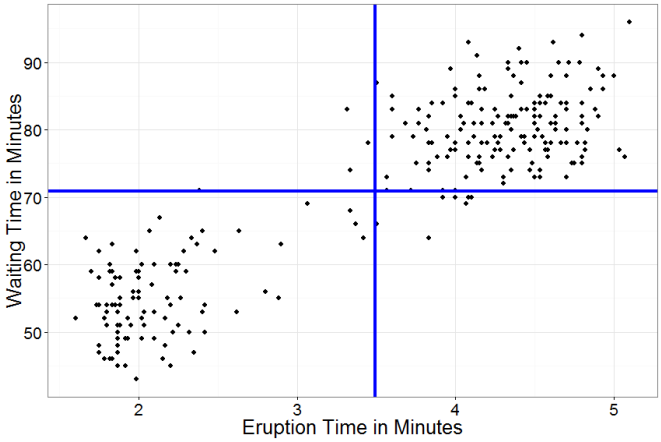


Figure 9: plot of chunk faithful2

Correlation Formula 2

- Formula for raw scores:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n)s_x s_y}$$

$$\rho = \frac{\sum(X - \mu_X)(Y - \mu_Y)}{(N)\sigma_x \sigma_y}$$

Correlation Properties

1. r ranges from -1 (perfect negative relationship) to +1 (perfect positive relationship)
2. $r = 0$ when there is no **linear** relationship between the variables.

3. The closer r is to +1 or -1, the stronger the relationship
    - For example, $r = -0.7$ is stronger than $r = 0.49$.
4. Changes in the scale are not uniform.
    - For example, a change from $r = 0.4$ does not represent half the relationship as $r = 0.8$

Correlation Properties 2

5. Linear transformations do not impact the correlation.
    - Example, the corelation between the temperatures in Boston and New York would be identical on Fahrenheit or Celsius scales

Figure 1. Daily maximum temperatures. New York and Boston, June 2005. The left hand panel plots the data in degrees Fahrenheit; the right hand panel, in degrees Celsius. This does not change r.

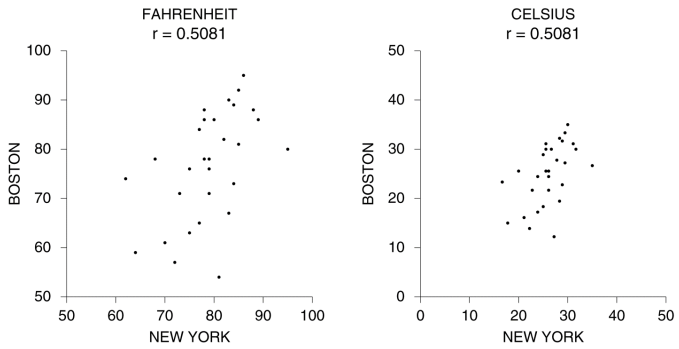


Figure 10:

6. Flipping the X and Y variables does not change the correlation.

Figure 2. Daily temperatures. New York, June 2005.

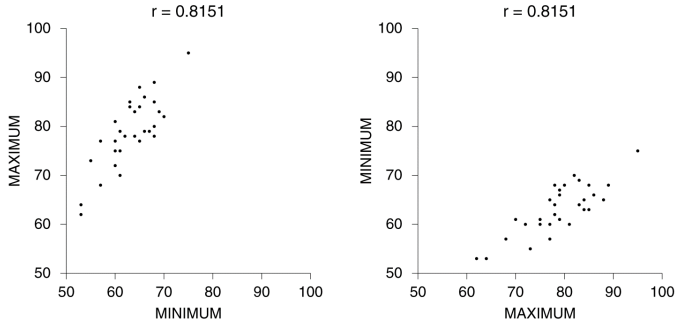


Figure 11:

Perfect Linearity

- Perfect linearity does not always imply perfect correlation.

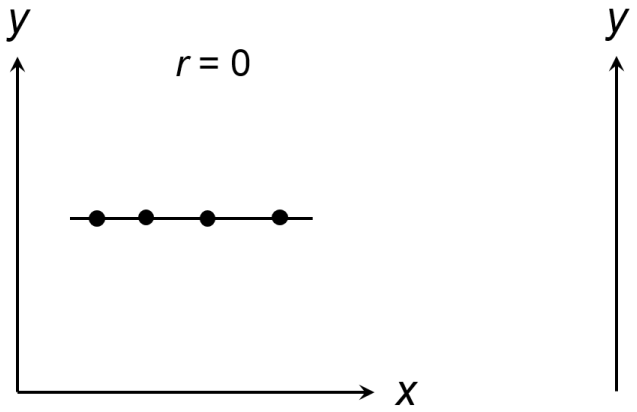


Figure 12:

Non-linear Trends

- Remember that the correlation we have discussed measures the **linear** relationship, not non-linear.

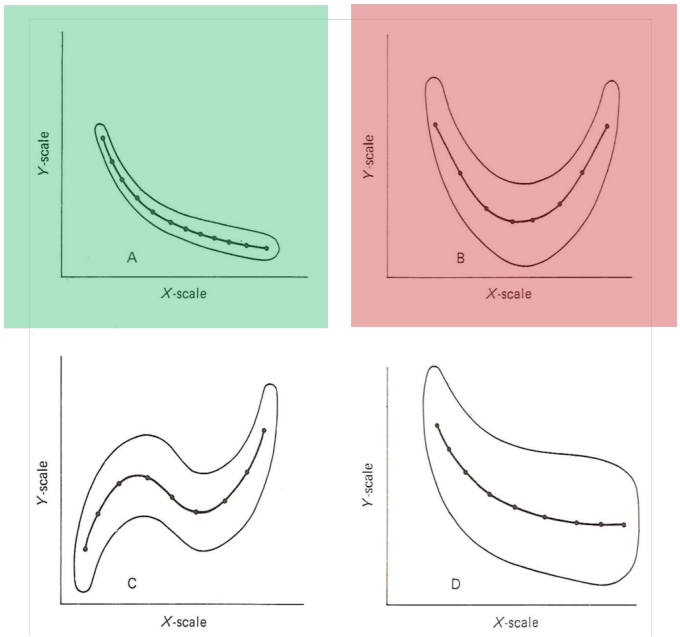


Figure 13:

Interpretation of the Correlation

- Assuming linearity exists, what does $r = 0.75$ mean?

1. Interpret r in an absolute sense (can be troublesome out of context):
   - $|r| > 0.75$; strong correlation
   - $|r| > 0.3$; medium correlation
   - $|r| < 0.3$; weak correlation
2. Interpret in a relative sense:
   - Parallel test forms (Form A – Form B); $r = 0.75$ is low
   - ACT to Freshman GPA; $r = 0.75$ is high

Interpretation of the Correlation 2

3. r^2 is an indication of the amount of variability one variable explains in the other variable.
   - Example: $r = 0.75$, $r^2 = 0.56$; 56% of the variation in Y is explained by X.
4. r as an index of prediction accuracy.

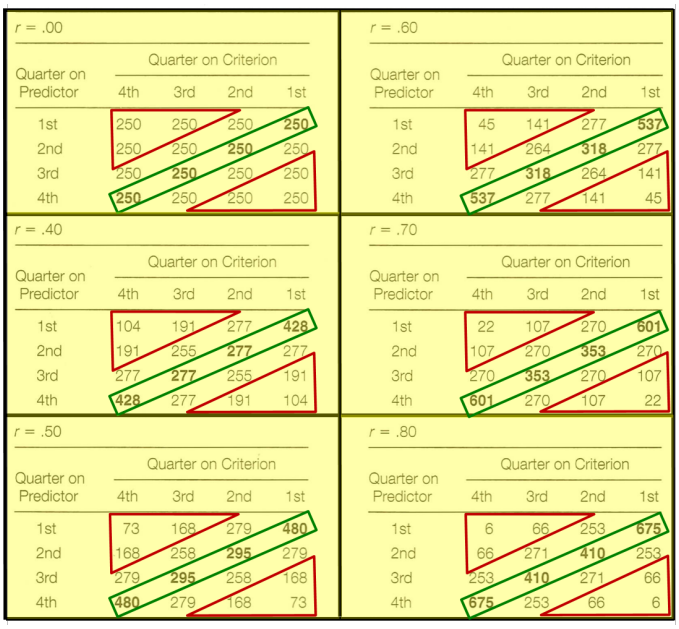


Figure 14:

Correlation is not Causation

- We find a relationship between X and Y

  - This relationship may be that X **causes** Y
  - Or it could be that Y **causes** X
  - It could be that a third variable causes both X and Y
- Most likely, there is a complex web of variables that are at play.

Confounding

- Confounding is when Y is caused by X and a third variable (but the third variable is not related to X). Therefore, the effect of X is confounded with a third variables

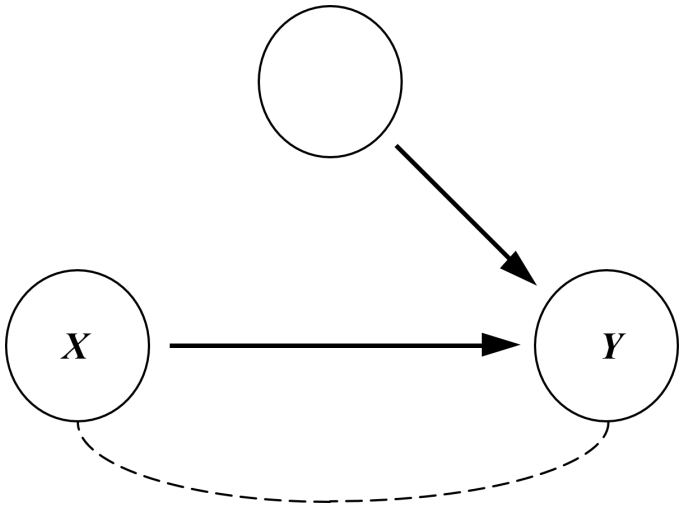


Figure 15:

Correlation with time

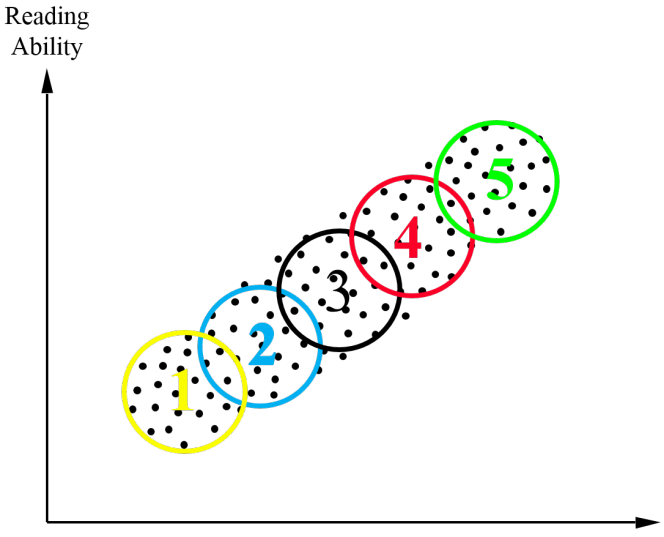


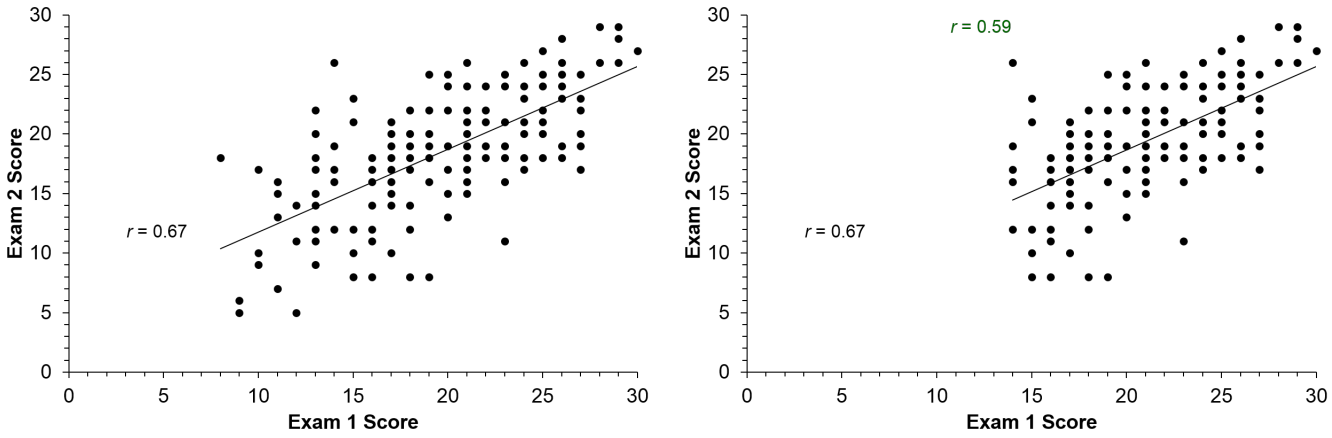
Figure 16:

Correlation is not Causation Summary

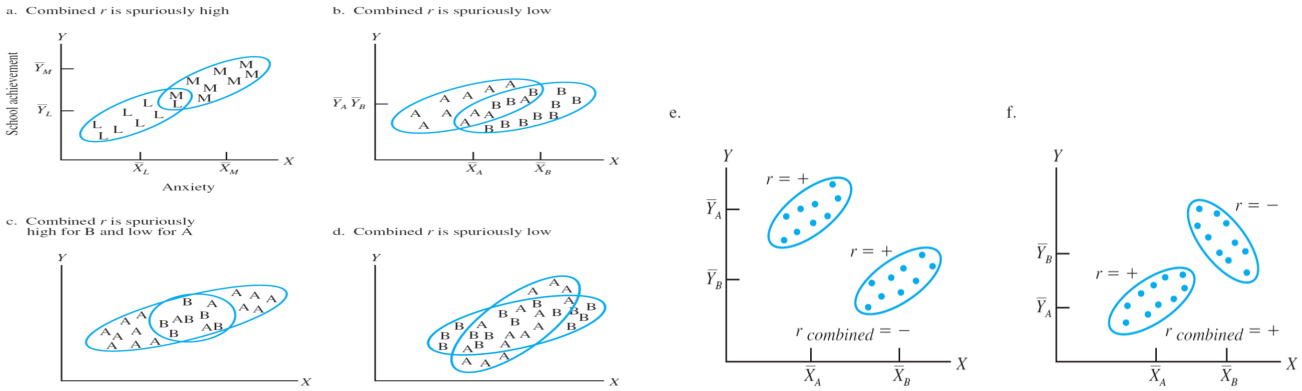
- Correlation specifically measures association/relationships
- Being associated or related is not the same as causation
- Inferences about causation can only be made with logic and careful experimental design/control.
- The value of r itself can not be used for causation

Restriction of Range

- Restriction of range can have the effect of decreasing the correlation.
- This happens due to decreasing the variability.



Correlations with subpopulations



Cautions for Correlation Summary

1. Careful about causation, remember correlation does not imply causation

2. Outliers can alter the correlation.
3. We discussed the correlation for **linear** relationships, not non-linear.
4. Restriction of range can decrease (attenuate) correlations.
5. More than one population can change the effect of the correlation.
6. Correlations of averages can be stronger than correlation of raw scores.