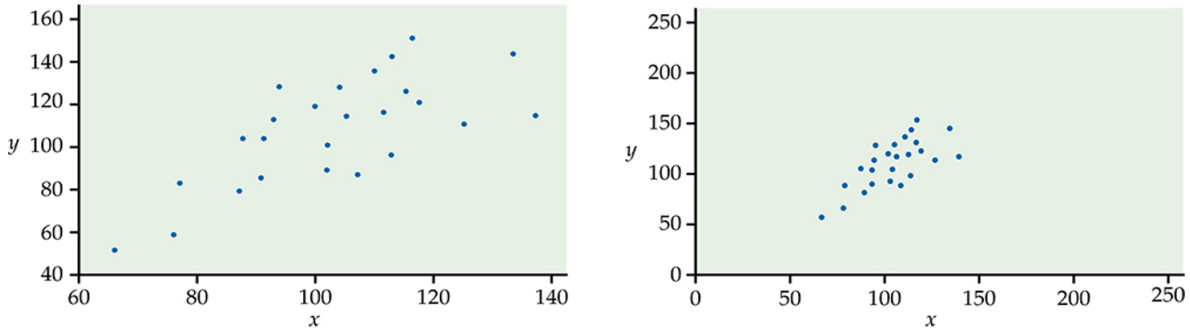
Topic 6: Correlation

Using Correlation to Measure Strength of Association

Which scatter plot below displays a stronger linear association?

These are the same data, plotted at different scales!



Moral: Visual inspection is not reliable for judging association strength.

Correlation, which we will denote with r, gives us a more mathematical way to measure the strength of the linear association between two numerical variables:

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

Requires (PAIRED) observations of 2 (numerical) variables

Data: n scatterplot-ready points (x_i, y_i)

(x_1, y_1), (x_2, y_2), ... , (x_n, y_n)

Preliminary computations:

Sample means of x AND y $\bar{x}$, $\bar{y}$

Sample standard deviations of x AND y s_x, s_y

Practice: Find correlation between X and Y given the observations below:

	X	Y
Obs. 1	3	4
Obs. 2	5	5
Obs. 3	7	3

Note: $n = 3$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}(3+5+7) = 5$ $\quad$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{3}(4+5+3) = 4$

Need sample VARIANCES first!

$s_x = \sqrt{s_x^2}$ $\quad$ $s_y = \sqrt{s_y^2}$

$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2}\left((3-5)^2 + (5-5)^2 + (7-5)^2\right) = \frac{8}{2} = 4$

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{2}\left((4-4)^2 + (5-4)^2 + (3-4)^2\right) = \frac{2}{2} = 1$

$s_x = \sqrt{4} = 2$ $\quad$ $s_y = \sqrt{1} = 1$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{3-1}\sum_{i=1}^{n}\left(\frac{x_i - 5}{2}\right)\left(\frac{y_i - 4}{1}\right)$$

$$= \frac{1}{2}\left[\left(\frac{3-5}{2}\right)\left(\frac{4-4}{1}\right) + \left(\frac{5-5}{2}\right)\left(\frac{5-4}{1}\right) + \left(\frac{7-5}{2}\right)\left(\frac{3-4}{1}\right)\right] = \frac{1}{2}[0+0+(-1)] = -\frac{1}{2}$$

Properties of Correlation

- Possible values — r is ALWAYS between -1 and 1, i.e. $-1 \le r \le 1$

  Value near ± 1 (i.e. near +1 or near -1) means STRONG association

  Value near 0 means WEAK association

- Sign meaning

  Positive sign (r > 0) = positive LINEAR relationship between X and Y

  Negative sign (r < 0) = negative LINEAR relationship between X and Y

- r does not change even if we...

  WARNING: we can ONLY use correlation meaningfully when examining LINEAR relationships

  ...swap which variable is X (explanatory) and which is Y (response)

  This change would alter scatterplot appearance and interpretation, but NOT the value of r!

- Note on outliers

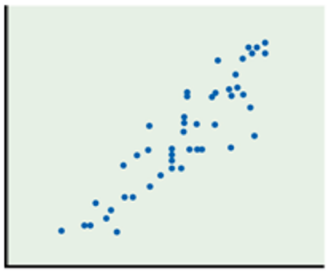
  r is NOT robust to outliers; it is SENSITIVE

  Outliers change r a lot, especially if we have only a small amount of data...

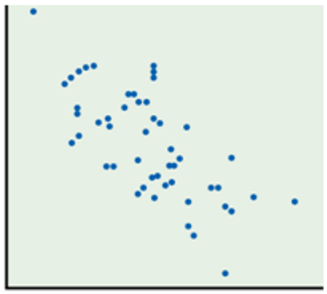
Examples:

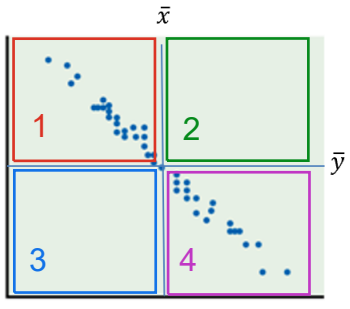

Correlation $r = 0$


Correlation $r = 0.5$


Correlation $r = 0.9$


Correlation $r = -0.3$


Correlation $r = -0.7$



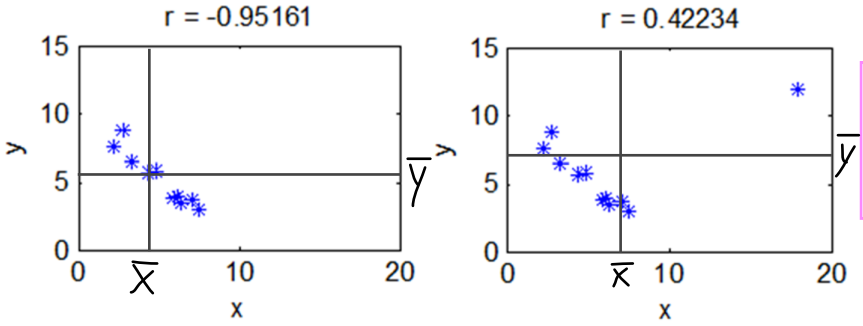
Correlation $r = -0.99$

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

	x-sign	y-sign	Product
1	-	+	-
2	+	+	+
3	-	-	+
4	+	-	-

How an outlier can distort an association:

(Disclaimer: sample mean lines are approximations)

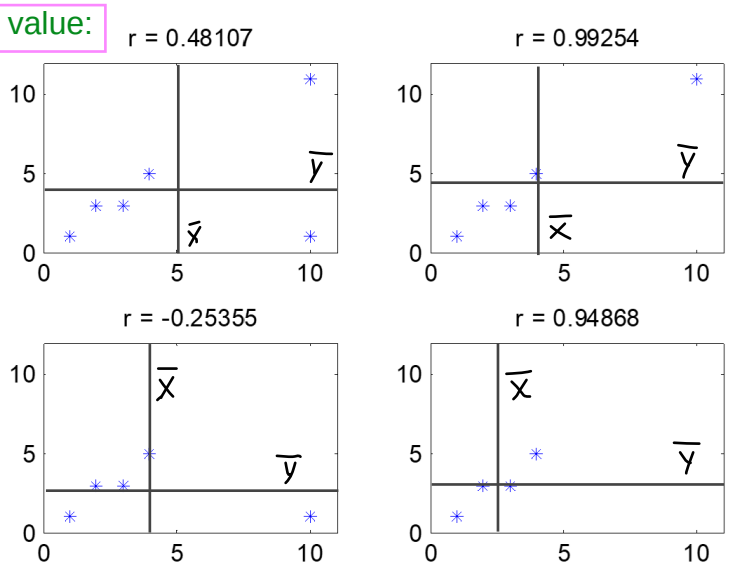


Turned a strong negative correlation into a weak positive one??

Another example:

Original value:

x	y
1	1
2	3
3	3
4	5
10	1
10	11

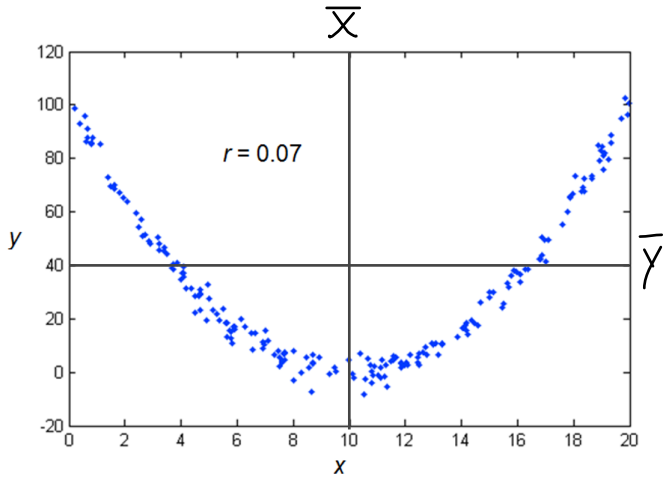


Looks stronger than the rest of the data would suggest

Without BOTH outliers: strong positive correlation

Looks negative and weak

Correlation does not describe nonlinear associations!

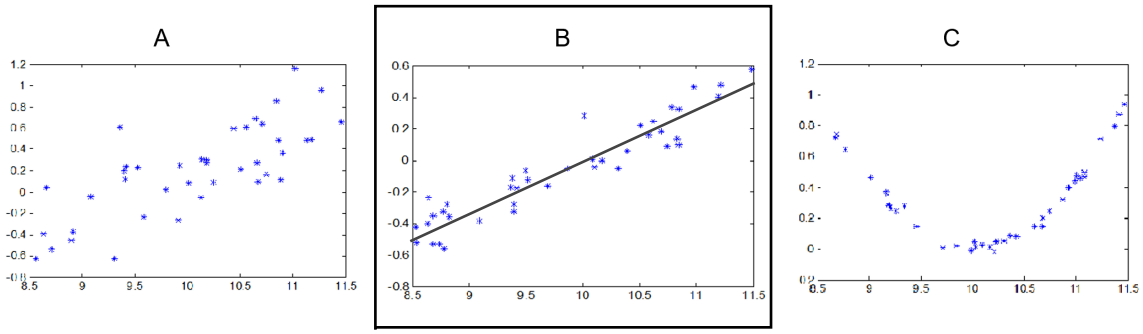


Quadrants balance out positive/negative contributions to correlation value!

Correlation according to r is weak, but association is clearly strong!

Moral: ALWAYS plot your data!

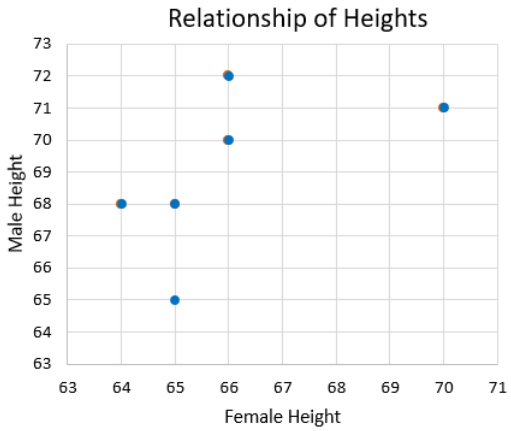
Practice: Which scatterplot has the largest correlation (closest to +1)?



Exercise: Heights of couples

This scatterplot (derived from the Excel sheet `heights.xlsx`) displays the paired heights of 6 heterosexual couples. On this plot, the value of r is:

Correct value: r = 0.565



1. If all the men were 6 inches shorter, would correlation change? Does the correlation tell us about whether women tend to date men taller than themselves?

Correlation would NOT change if all y-values were smaller by 6

We can say: As women's height increases, men's height TENDS TO increase as well

"Correlation is **invariant under translation**"

Note: NOT a causative statement!

2. If heights were in centimeters, would correlation change?

Correlation would NOT change if heights were in centimeters (i.e. if heights were "rescaled" by some fixed ratio)

"Correlation is **invariant under scaling**"

3. If each woman dated a man exactly 3 in. taller than herself, what would be the correlation?

In this case, **r = 1**

We can write an equation for a STRAIGHT LINE that all of the data would exactly land on!

r closer to ± 1 means data is closer to forming some straight line

M = F + 3