STAT 22000 Lecture Slides
Correlation

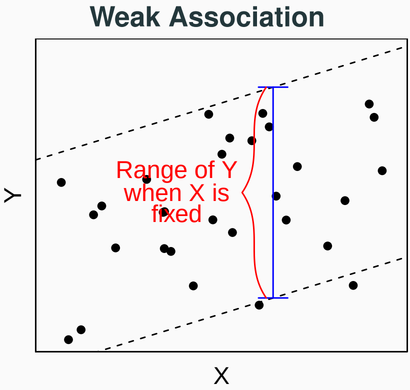
Yibi Huang
Department of Statistics
University of Chicago

Coverage

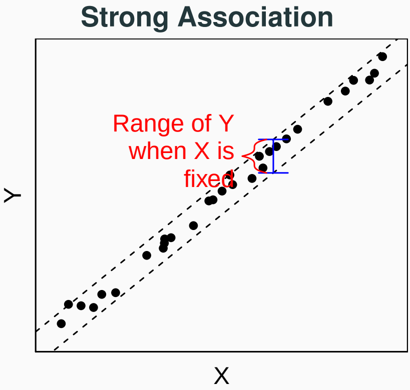
This set of slides covers Section 8.1.4 in the 4th edition of *OpenIntro Statistics*[1] and more.

[1] or Section 7.1.4 in the 3rd edition

Recall when we introduced scatter plots in Chapter 1, we assessed the **strength** of the association between two variables by eyeballs.



Large spread of Y when X is known



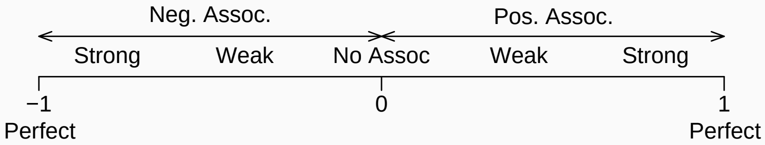
Small spread of Y when X is known

Correlation = Correlation Coefficient, r

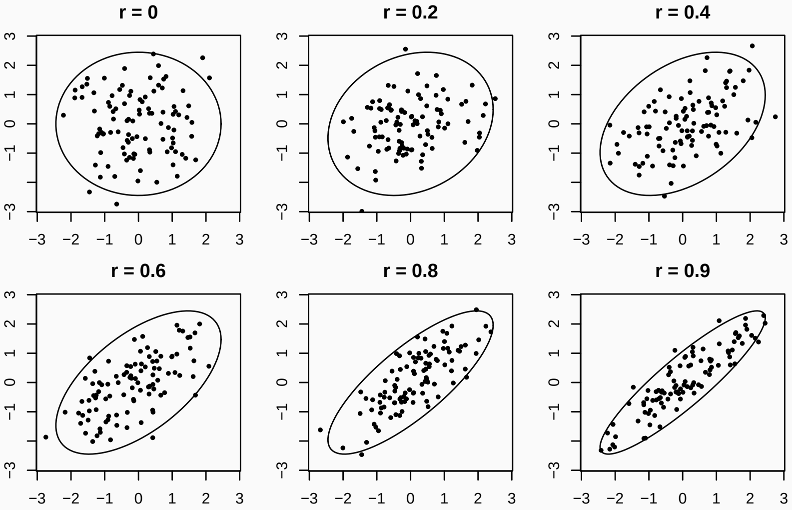
Correlation r is a numerical measure of the *direction* and *strength* of the **linear** relationship between two numerical variables.

"r" always lies between -1 and 1; the strength increases as you move away from 0 to either -1 or 1.

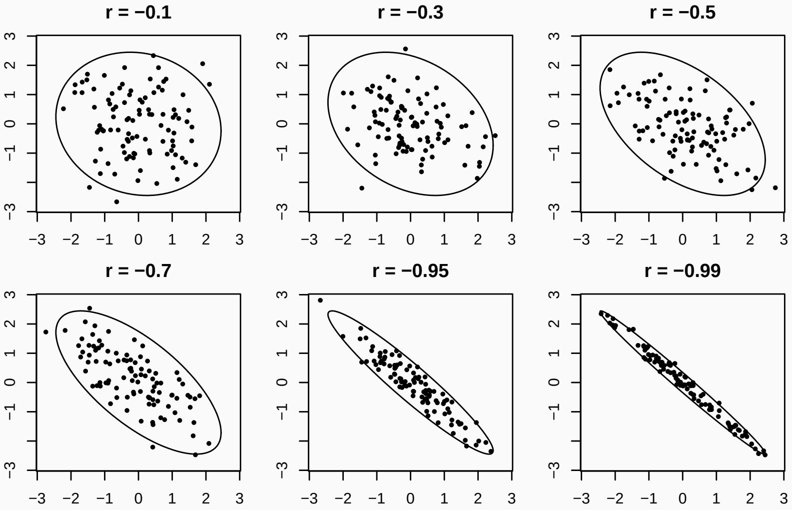
- $r > 0$: positive association
- $r < 0$: negative association
- $r \approx 0$: very weak linear relationship
- large $|r|$: strong linear relationship
- $r = -1$ or $r = 1$: *only* when all the data points on the scatterplot lie exactly along a *straight line*



Positive Correlations



Negative Correlations



Formula for Computing the Correlation Coefficient "r"

(x_1, y_1)

(x_2, y_2)

(x_3, y_3)

$\vdots$

(x_n, y_n)

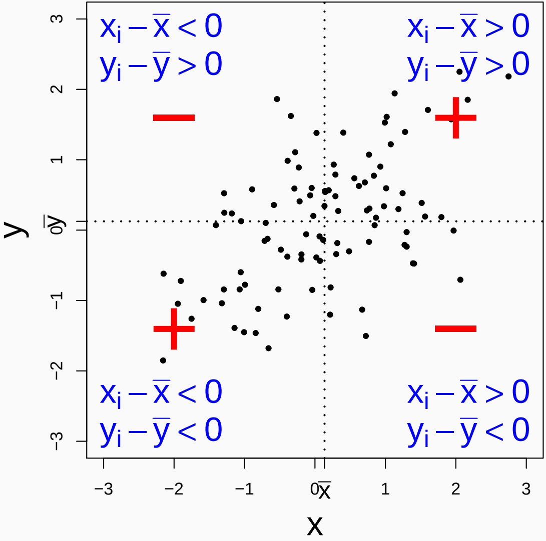
The **correlation coefficient** r (or simply, **correlation**) is defined as:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z\text{-score of } x_i} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z\text{-score of } y_i}.$$

where s_x and s_y are respectively the sample SD of X and of Y.

Usually, we find the correlation using softwares rather than by manual computation.

Why r Measures the Strength of a Linear Relationship?



What is the sign of $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$??

Here $r > 0$; more positive contributions than negative.

What kind of points have large contributions to the correlation?

Correlation r Has No Unit

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z\text{-score of } x_i} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z\text{-score of } y_i}.$$

After standardization, the z-score of neither x_i nor y_i has a unit.

- So r is unit-free.
- So we can compare r between data sets, where variables are measured in different units or when variables are different. E.g. we may compare the

  r between [swim time and pulse],

  with the

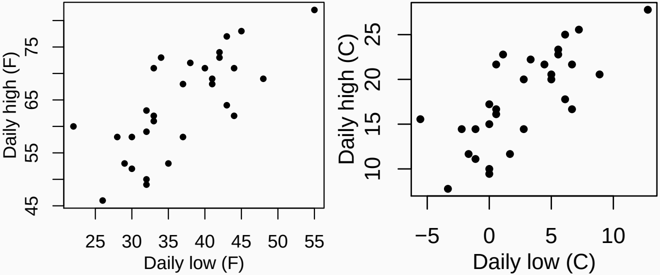
  r between [swim time and breathing rate].

Correlation r Has No Unit (2)

Changing the units of variables does not change the correlation coefficient r, because we get rid of all the units when we standardize them (get z-scores).

E.g., no matter the temperatures are recorded in $^\circ F$, or $^\circ C$, the correlations obtained are equal because

$$C = \frac{5}{9}(F - 32).$$

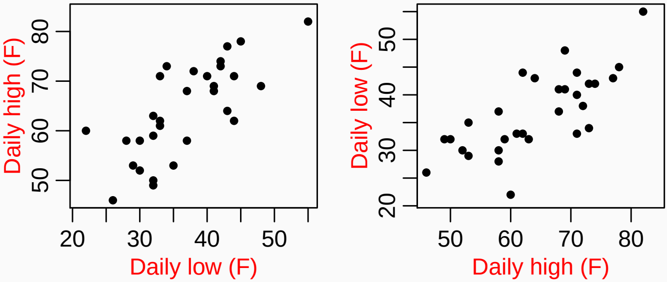


"r" Does Not Distinguish x & y

Sometimes one use the X variable to predict the Y variable. In this case, X is called the *explanatory variable*, and Y the *response*. The correlation coefficient r does not distinguish between the two. It treats x and y symmetrically.

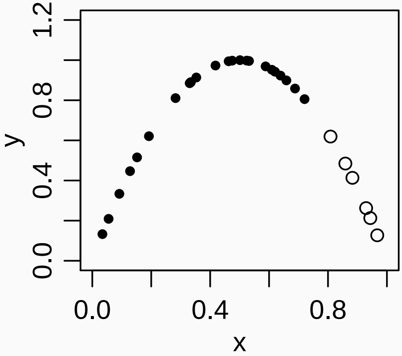
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Swapping the x-, y-axes doesn't change r (both $r = 0.74$.)



Correlation r Describes Linear Relationships Only

The scatter plot below shows a *perfect nonlinear* association. All points fall on the quadratic curve $y = 1 - 4(x - 0.5)^2$.



r of all black dots $= 0.803$,

r of all dots $= -0.019$.
(black + white)

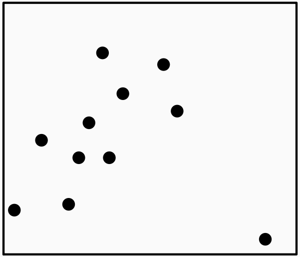
No matter how strong the association,

the r of a curved relationship is NEVER 1 or –1.

It can even be 0, like the plot above.

Correlation Is VERY Sensitive to Outliers

Sometimes a single outlier can change r drastically.



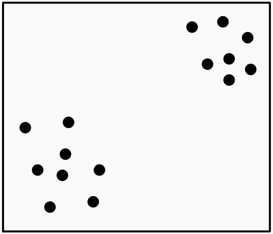
For the plot on the left,

$$r = \begin{cases} 0.0031 & \text{with the outlier} \\ 0.6895 & \text{without the outlier} \end{cases}$$

Outliers that may remarkably change the form of associations when removed are called **influential points**.

Remark: Not all outliers are influential points.

When Data Points Are Clustered ...



In the plot above, each of the two clusters exhibits a weak negative association ($r = -0.336$ and -0.323).

But the whole diagram shows a moderately strong positive association ($r = 0.849$).

- This is an example of the **Simpson's paradox**.
- An overall r can be misleading when data points are clustered.
- Cluster-wise r's should be reported as well.

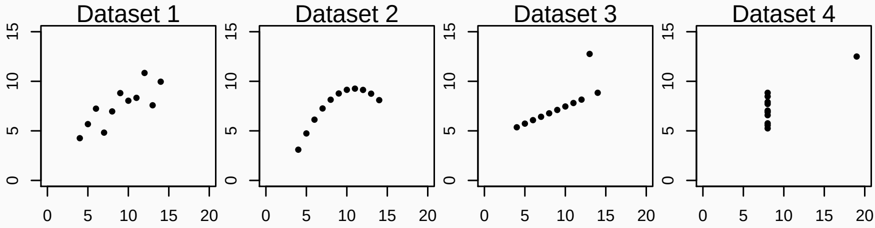
Always Check the Scatter Plots (1)

The 4 data sets below have identical $\overline{x}$, $\overline{y}$, s_x, s_y, and r.

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.96	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.75	13	12.76	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.36	19	12.50
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Ave	9	7.5	9	7.5	9	7.5	9	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

How about their scatter plots?

Always Check the Scatter Plots (2)



- In Dataset 2, y can be predicted exactly from x. But $r < 1$, because r only measures **linear** association.
- In Dataset 3, r would be 1 instead of 0.82 if the outlier were actually on the line.

The correlation coefficient can be misleading in the presence of outliers, multiple clusters, or nonlinear association.

Correlation Indicates Association, *Not* Causation

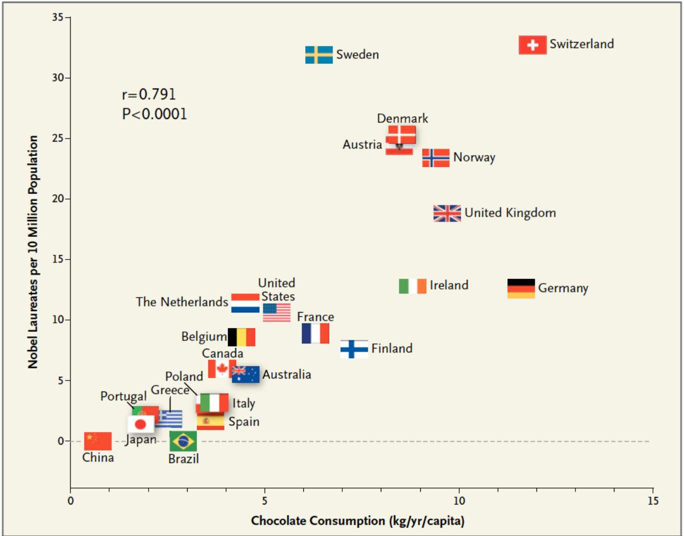


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: http://www.nejm.org/doi/full/10.1056/NEJMon1211064

Questions

- Why do both variables have to be numerical when computing their correlation coefficient?
- If the law requires women to marry only men 2 years older than themselves, what is the correlation of the ages between husbands and wives?

Husbands' age = Wife's age + 2
All the points would fall on the line y = x + 2.
r = 1 or -1. r must be 1 since the slope is positive.