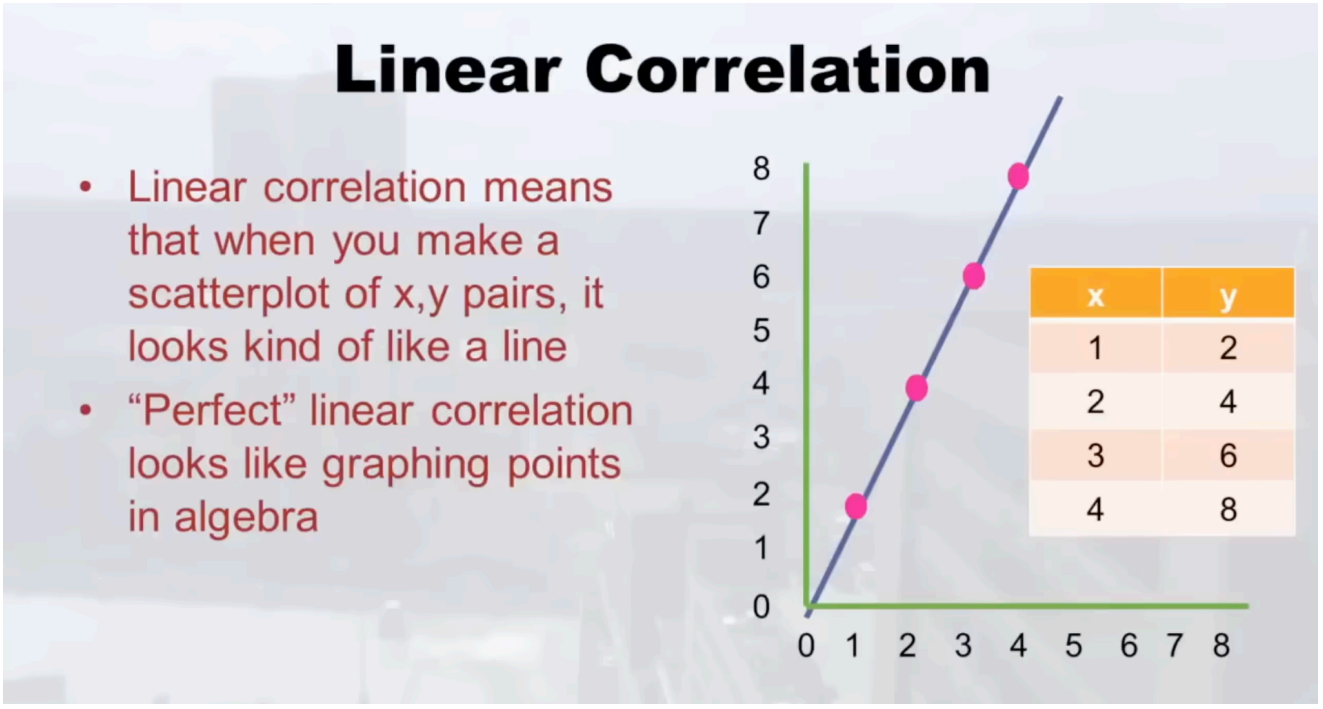
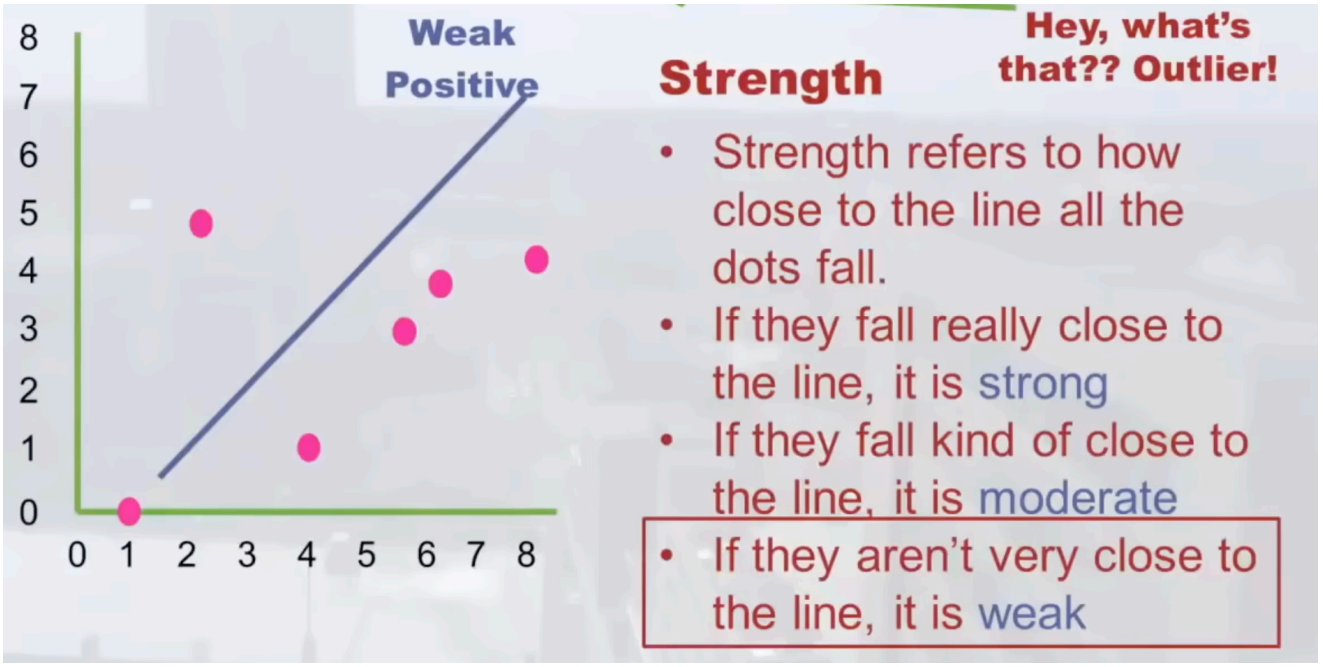
4.1 Scatter Diagram and Linear correlation

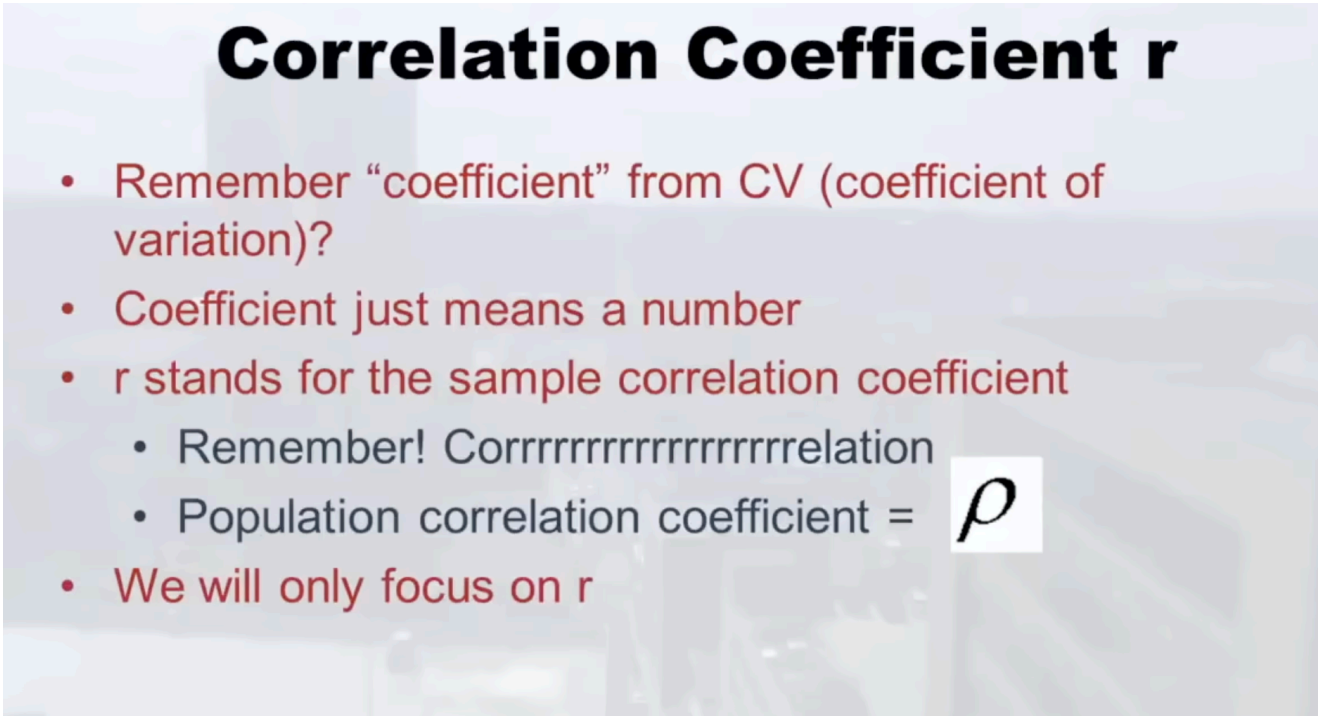
Scattergram = Scatter plot
-Explanatory (Independent)
Variable is called X
-Graphed on x-axis
-Response (dependent)
Variable is called Y
-Graphed on y-axis

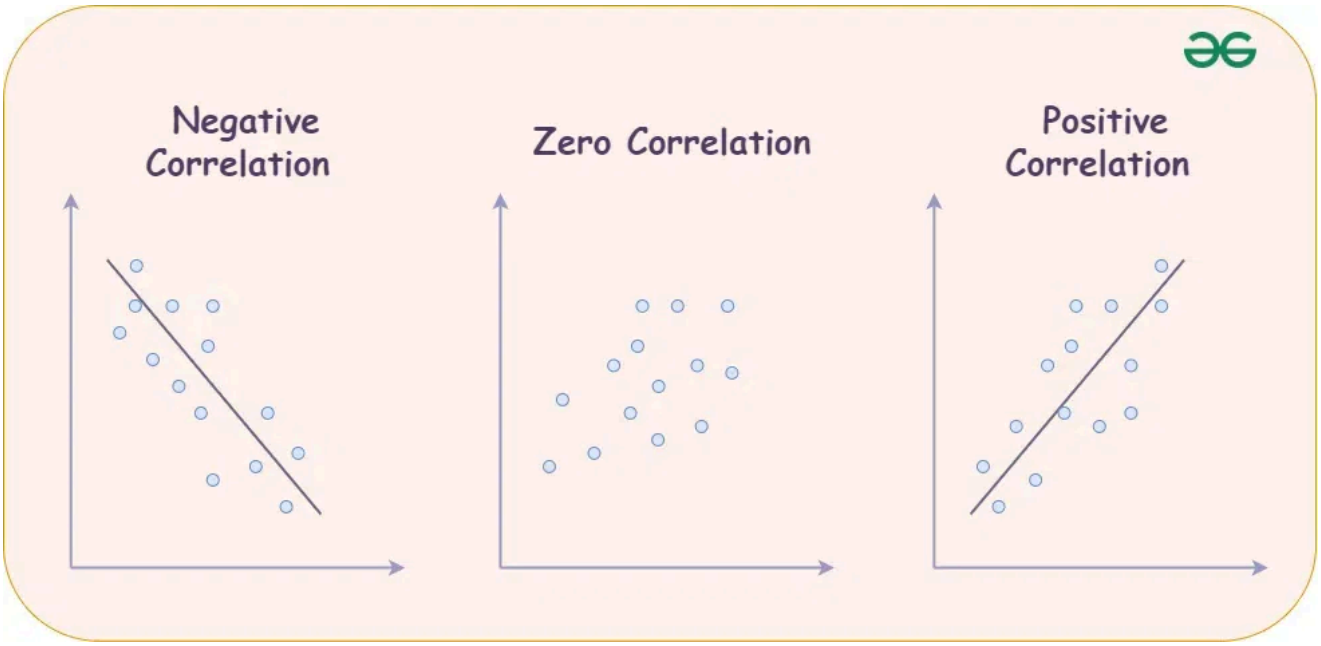


Correlation has two attribute
-Positive correlation
-Negative correlation
-Moderate correlation



Correlation Coefficient r





What is r?

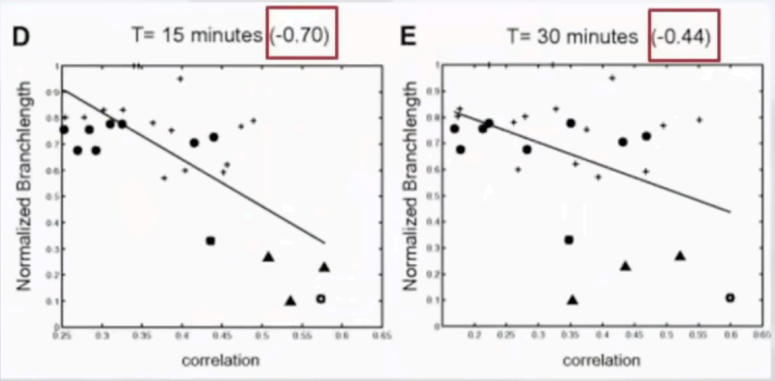
What it is

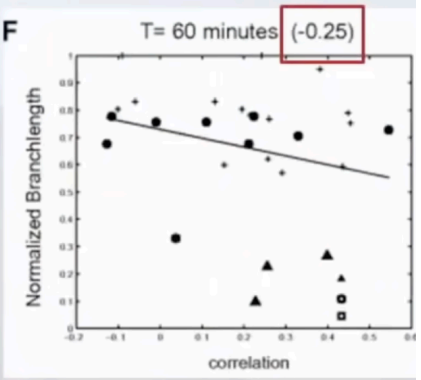
- A numerical quantification of how correlated a set of x,y pairs are
- Calculated from plugging x,y pairs into an equation
- Has a defining formula and a computational formula
- I will demonstrate computational formula

How to interpret it

- The r calculation produces a number
- The lowest number possible is -1.0
  - Perfect negative correlation
- The highest possible number is 1.0
  - Perfect positive correlation
- All others are in-between

Examples of Negative r





r = -0.70 **r = -0.44**

Evolutionary principles of modular gene regulation in yeasts

Figure 11.

Dawn A Thompson, Sushmita Roy, Michelle Chan, Mark P Styczynsky, Jenna Pfiffner, Courtney French, Amanda Socha, Anne Thielke, Sara Napolitano, Paul Muller, Manolis Kellis, Jay H Konieczka, Ilan Wapinski, Aviv Regev

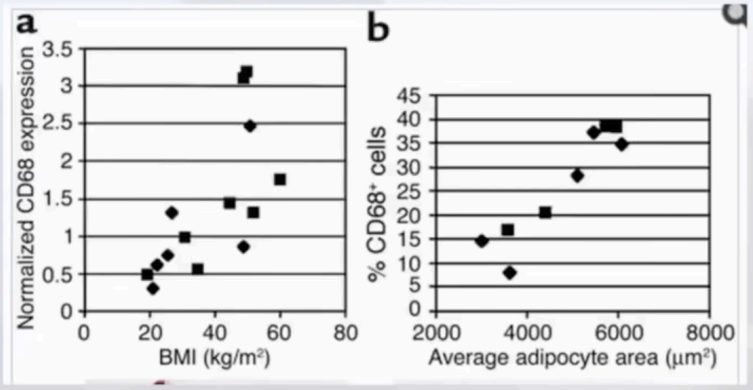
Broad Institute of MIT and Harvard, United States; Massachusetts Institute of Technology, United States; Howard Hughes Medical Institute, Massachusetts Institute of Technology, United States

OPINION!!! **r = -0.25**

For negative correlations:

- 0.0 to -0.40: Weak
- -0.40 to -0.70: Moderate
- -0.70 to -1.0: Strong

Examples of Positive r



r = 0.66 **r = 0.92**

Obesity is associated with macrophage accumulation in adipose tissue

Stuart P. Weisberg,[1] Daniel McCann,[1] Manisha Desai,[2] Michael Rosenbaum,[1] Rudolph L. Leibel,[1,3,4] and Anthony W. Ferrante, Jr.[3,4]

OPINION!!!

For positive correlations:

- 0.0 to 0.40: Weak
- 0.40 to 0.70: Moderate
- 0.70 to 1.0: Strong

- -1 indicates a strong negative relationship
- 1 indicates strong positive relationships
- Zero implies no connection at all

Computational Formula

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Hypothetical Scenario

- We have 7 patients
- They have come to the clinic for appointments throughout the year.
- We predict those with a higher diastolic blood pressure (DBP) will have more appointments
- We take DBP at last appointment as "x"
- We take number of appointments over the year as "y"

- *FLASHBACK!* …to Chapter 3.2
- Notice all the Σ's
- As before, we will
  - make columns
  - make calculations
  - Then add up the columns to get these Σ's

x=DBP, y=# of Appointments

#	x	y	x^2	y^2	xy
1	70	3	4,900	9	210
2	115	45	13,225	2,025	5,175
3	105	21	11,025	441	2,205
4	82	7	6,724	49	574
5	93	16	8,649	256	1,488
6	125	62	15,625	3,844	7,750
7	88	12	7,744	144	1,056
	Σx = 678	Σy = 166	Σx^2 = 67,892	Σy^2 = 6,768	Σxy = 18,458

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$\frac{(7)(18,458) - (678)(166)}{\sqrt{(7)(67,892) - (678)^2}\sqrt{(7)(6,768) - (166)^2}}$$

$$\frac{16,658}{124.74 * 140.78 = 17,561.3}$$

$$r = \frac{16,658}{17,561.3} = 0.949$$

OPINION! 0.70 to 1.0: Strong

Facts About r

- r requires data with a "bivariate normal distribution" – we do not cover looking at this in this class, but please know this.
- r does not have units.
- Perfect linear correlation is r=-1.0 or r=1.0 (depending on direction). No linear correlation is r=0.
- Positive r means as x goes up, y goes up, and as x goes down, y goes down.
- Negative r means as x goes up, y goes down, and as x goes down, y goes up.
- Even if you switched x and y on the axes, you'd get the same r.
- Even if you converted x and y to different units (e.g., you converted measurements into the metric system), you'd get the same r.

Lurking Variables and 'Correlation is not Causation'

Examples

Claim

- Over time, as people purchase more onions, the stock market rises. This is true for many generations in the US.

Reality

- "A healthy economy" is the lurking variable
  - A healthy economy makes people be able to afford more food (including onions).
  - A healthy economy boosts the stock market.