Statistical reasoning for humanities scholars

Lecture III: covariance

Nalanda Academy Workshop 17-19 May 2025

Subodh Patil and Kapil Wankhede

Standard deviation = square root of variance

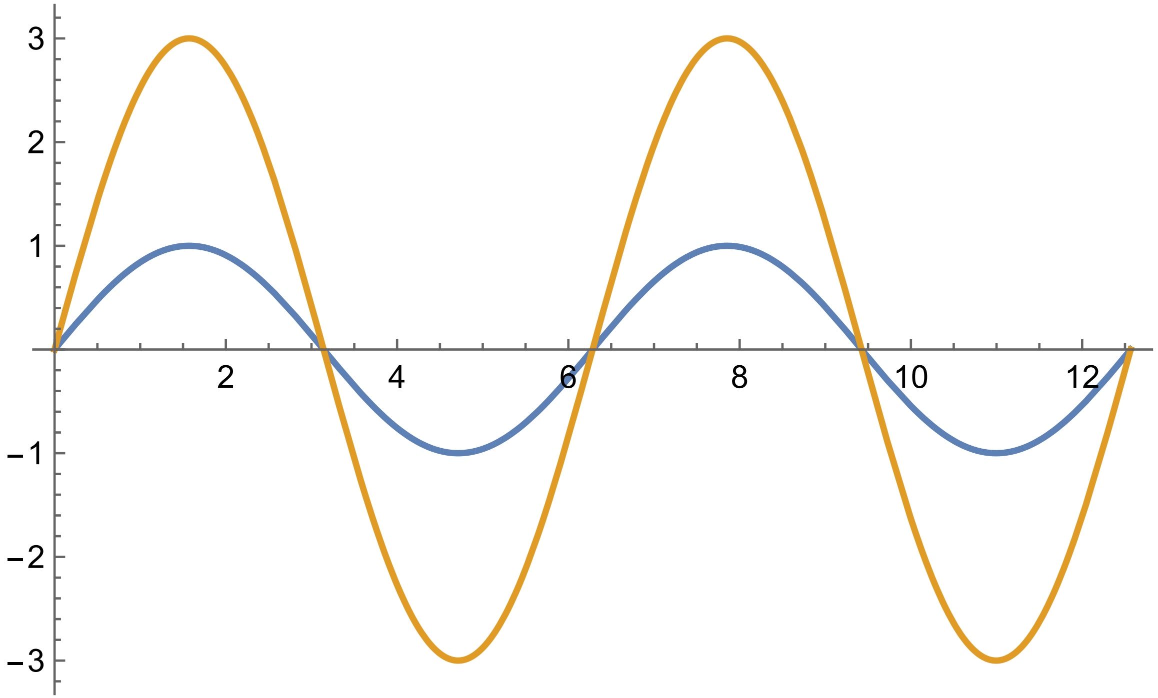
- Variance = average of the deviation squared:

$$\langle (x - \langle x \rangle)^2 \rangle = \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)$$

- Standard deviation =

$$\sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)}$$

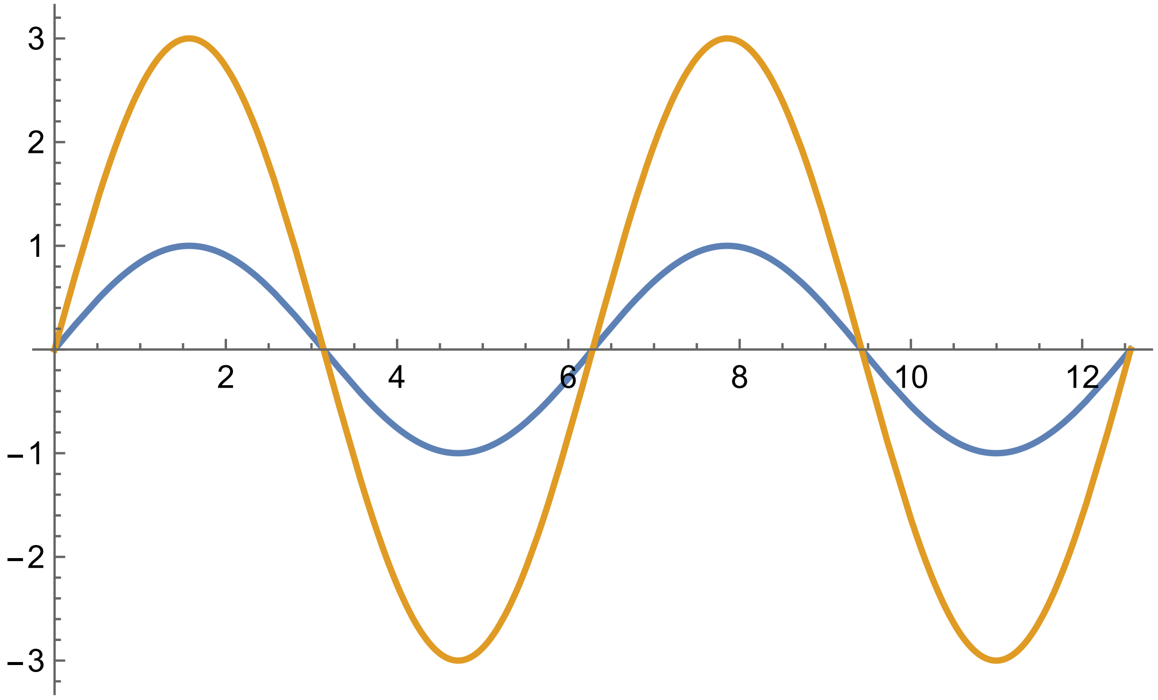
Consider AC voltage:



Average of both signals vanish, but standard deviation is different:

$$\sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)}$$

Consider AC voltage:



Standard deviation of voltage = root mean squared (RMS) voltage

$$\sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)}$$

Average of 1000 consecutive coin tosses:

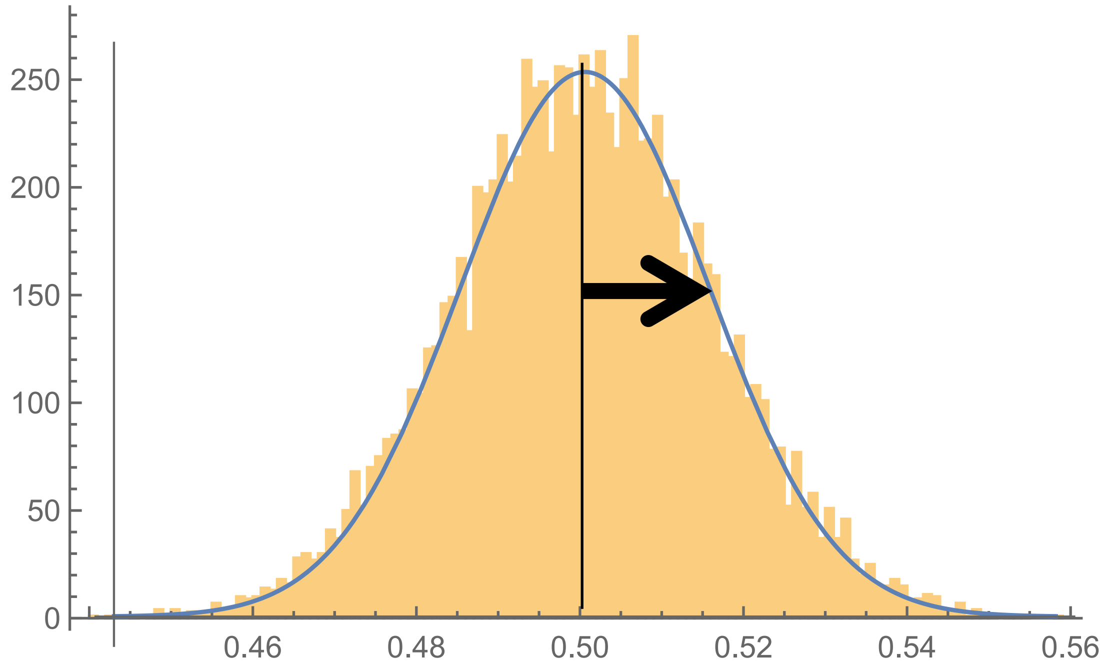
- Average = 0.5

- Standard deviation:

$$\sqrt{\langle (x - \langle x \rangle)^2 \rangle} = \sqrt{\sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)}$$

$$= \frac{1}{2\sqrt{N}}$$

$$= 0.0158$$



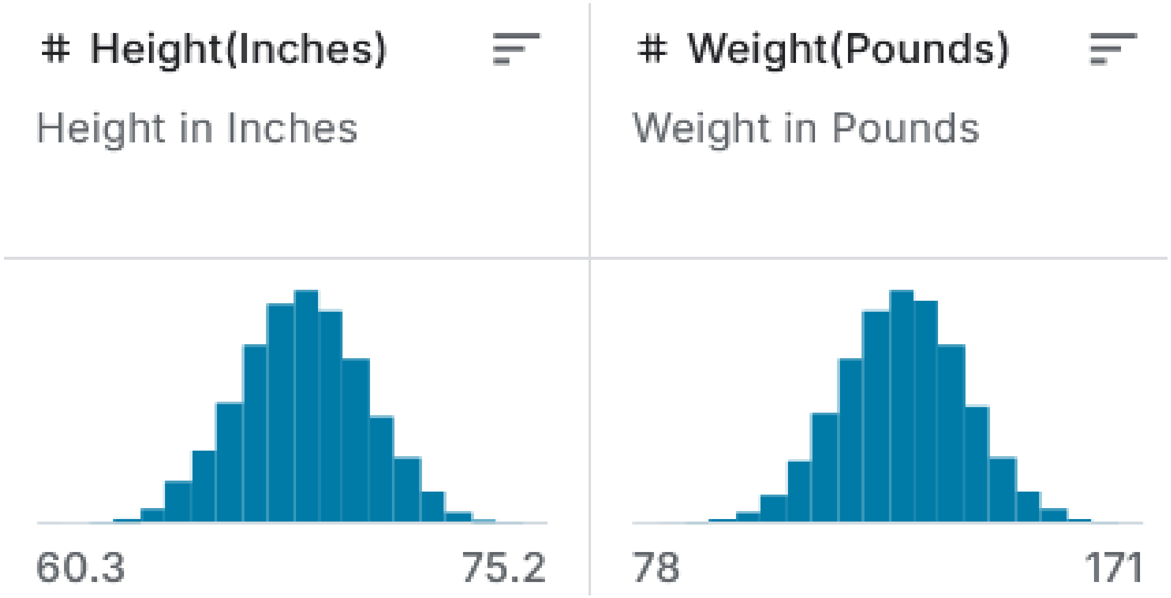
Variance and *co*variance

- Variance = average of the deviation squared:

$$\langle (x - \langle x \rangle)^2 \rangle = \sum_{i=1}^{N} (x_i - \langle x \rangle)^2 p(x_i)$$

- But what about possible relations between one random variable and another?

Variance and covariance

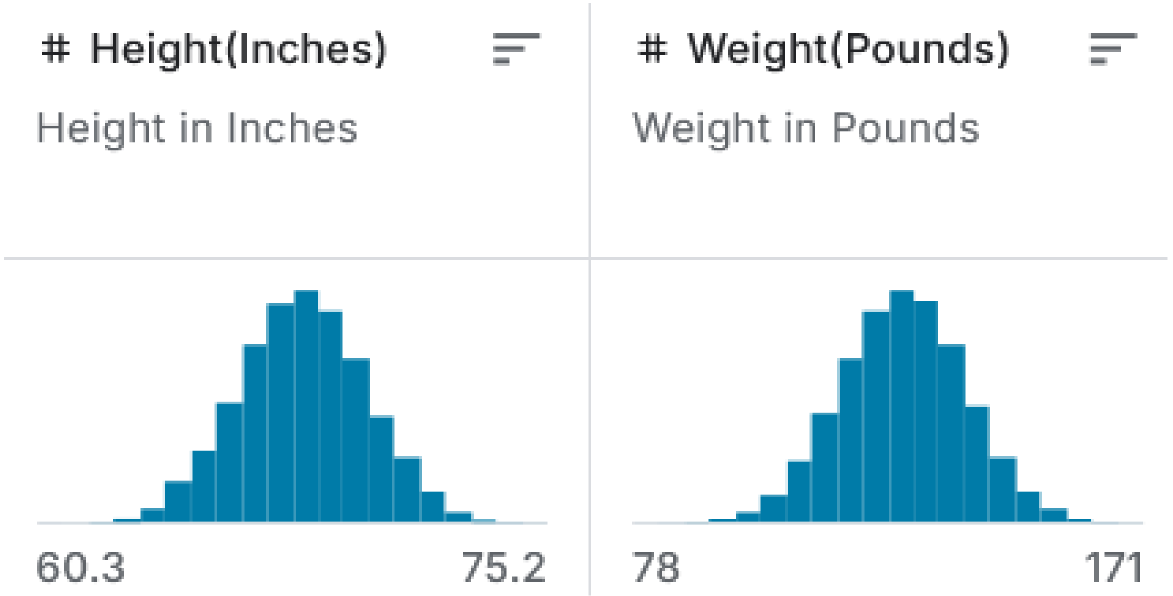


# Index	# Height(Inc...	# Weight(Po...
1	65.78331	112.9925
2	71.51521	136.4873
3	69.39874	153.0269
4	68.2166	142.3354
5	67.78781	144.2971
6	68.69784	123.3024
7	69.80204	141.4947
8	70.01472	136.4623
9	67.90265	112.3723
10	66.78236	120.6672
11	66.48769	127.4516
12	67.62333	114.143
13	68.30248	125.6107
14	67.11656	122.4618
15	68.27967	116.0866

Data of heights and weights of N = 25K people

https://www.kaggle.com/datasets/burnoutminer/heights-and-weights-dataset

Variance and covariance



# Index	# Height(Inc...	# Weight(Po...
1	65.78331	112.9925
2	71.51521	136.4873
3	69.39874	153.0269
4	68.2166	142.3354
5	67.78781	144.2971
6	68.69784	123.3024
7	69.80204	141.4947
8	70.01472	136.4623
9	67.90265	112.3723
10	66.78236	120.6672
11	66.48769	127.4516
12	67.62333	114.143
13	68.30248	125.6107
14	67.11656	122.4618
15	68.27967	116.0866

Taller than average people tend to be heavier than average, shorter than average people tend to be lighter than average.

Variance and covariance

Taller than average people tend to be heavier than average, shorter than average people tend to be lighter than average.

If (height – average height) > 0, then typically also (weight – average weight) > 0

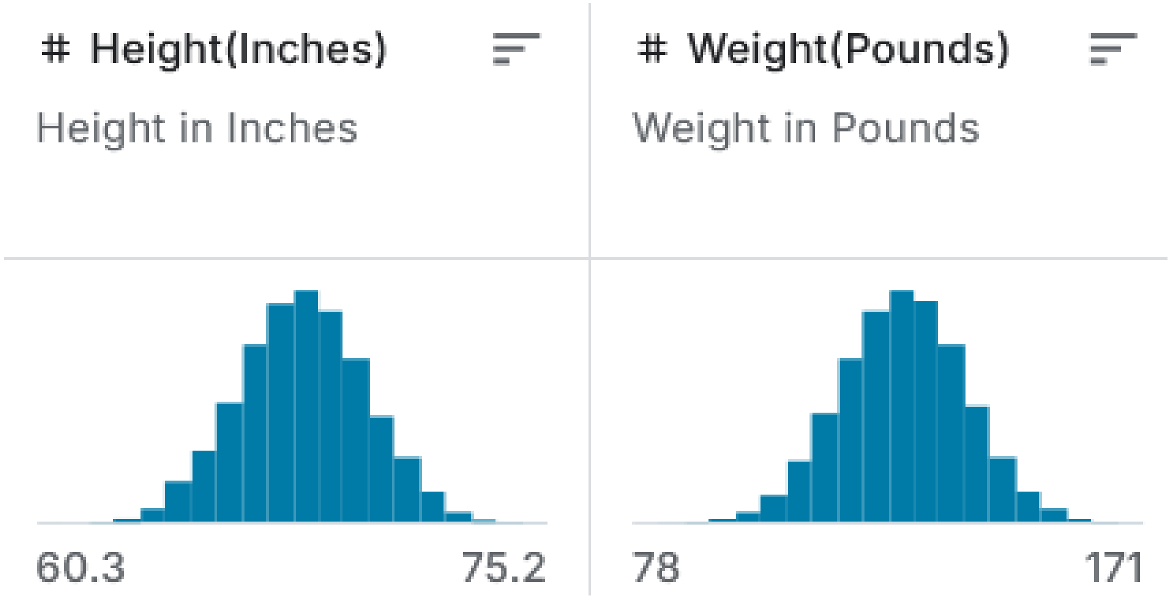
… but also

If (height – average height) < 0, then typically also (weight – average weight) < 0

Therefore try:

$$\langle (h - \langle h \rangle)(w - \langle w \rangle) \rangle = \frac{1}{N} \sum_i^N (h_i - \langle h \rangle)(w_i - \langle w \rangle) \equiv Cov(h, w)$$

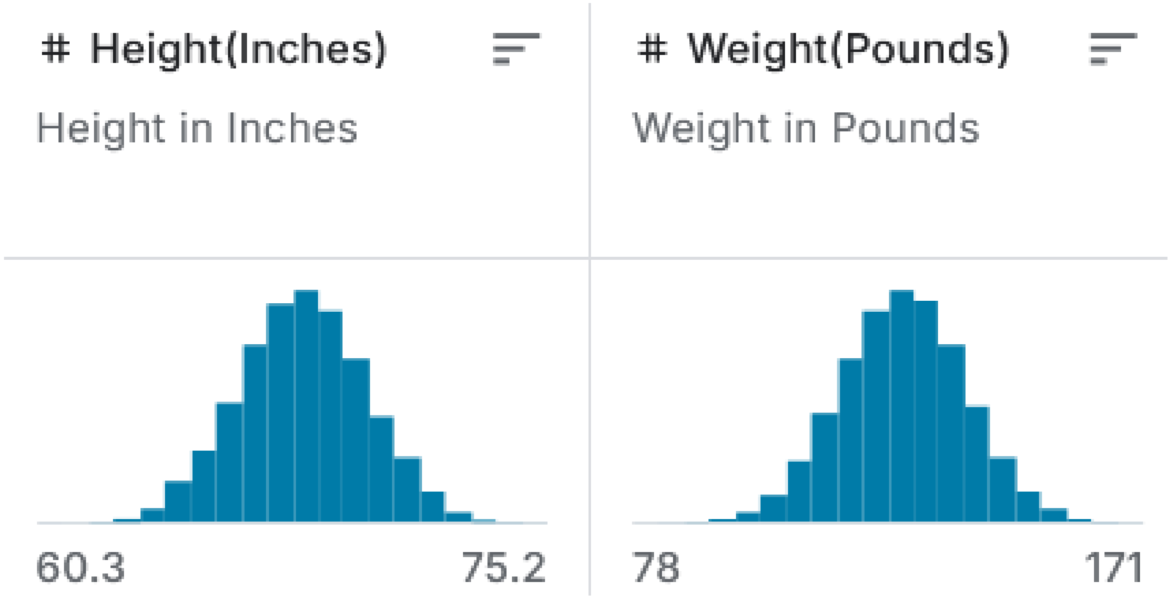
Variance and covariance



# Index	# Height(Inc...	# Weight(Po...
1	65.78331	112.9925
2	71.51521	136.4873
3	69.39874	153.0269
4	68.2166	142.3354
5	67.78781	144.2971
6	68.69784	123.3024
7	69.80204	141.4947
8	70.01472	136.4623
9	67.90265	112.3723
10	66.78236	120.6672
11	66.48769	127.4516
12	67.62333	114.143
13	68.30248	125.6107
14	67.11656	122.4618
15	68.27967	116.0866

But taller than average people may or may not be smarter than on average. Height has nothing to do with intelligence!

Variance and covariance



# Index	# Height(Inc...	# Weight(Po...
1	65.78331	112.9925
2	71.51521	136.4873
3	69.39874	153.0269
4	68.2166	142.3354
5	67.78781	144.2971
6	68.69784	123.3024
7	69.80204	141.4947
8	70.01472	136.4623
9	67.90265	112.3723
10	66.78236	120.6672
11	66.48769	127.4516
12	67.62333	114.143
13	68.30248	125.6107
14	67.11656	122.4618
15	68.27967	116.0866

Let e_i denote exam score of i'th person:

$$\langle (h - \langle h \rangle)(e - \langle e \rangle) \rangle = \frac{1}{N} \sum_i^N (h_i - \langle h \rangle)(e_i - \langle e \rangle) \equiv Cov(h,e) \approx 0$$

Variance and *co*variance

If Cov(x,y) > 0, we say x and y are *positively correlated* (e.g. height and weight)

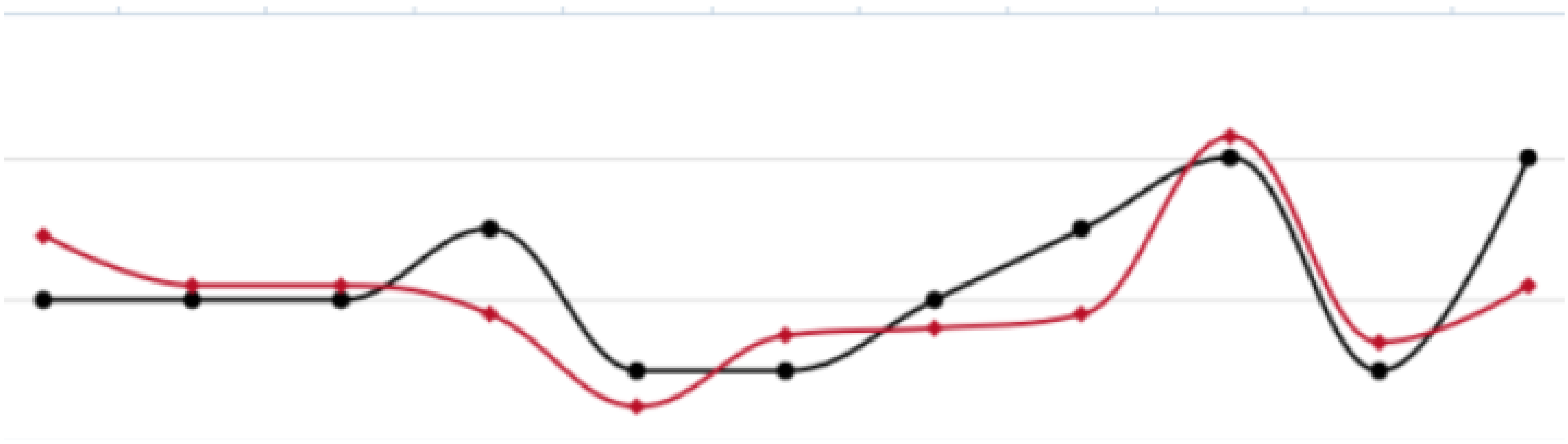
If Cov(x,y) < 0, we say x and y are *negatively correlated* (e.g. PCI and fertility rates)

If Cov(x,y) = 0, we say x and y are *uncorrelated* (e.g. height and math exam scores)

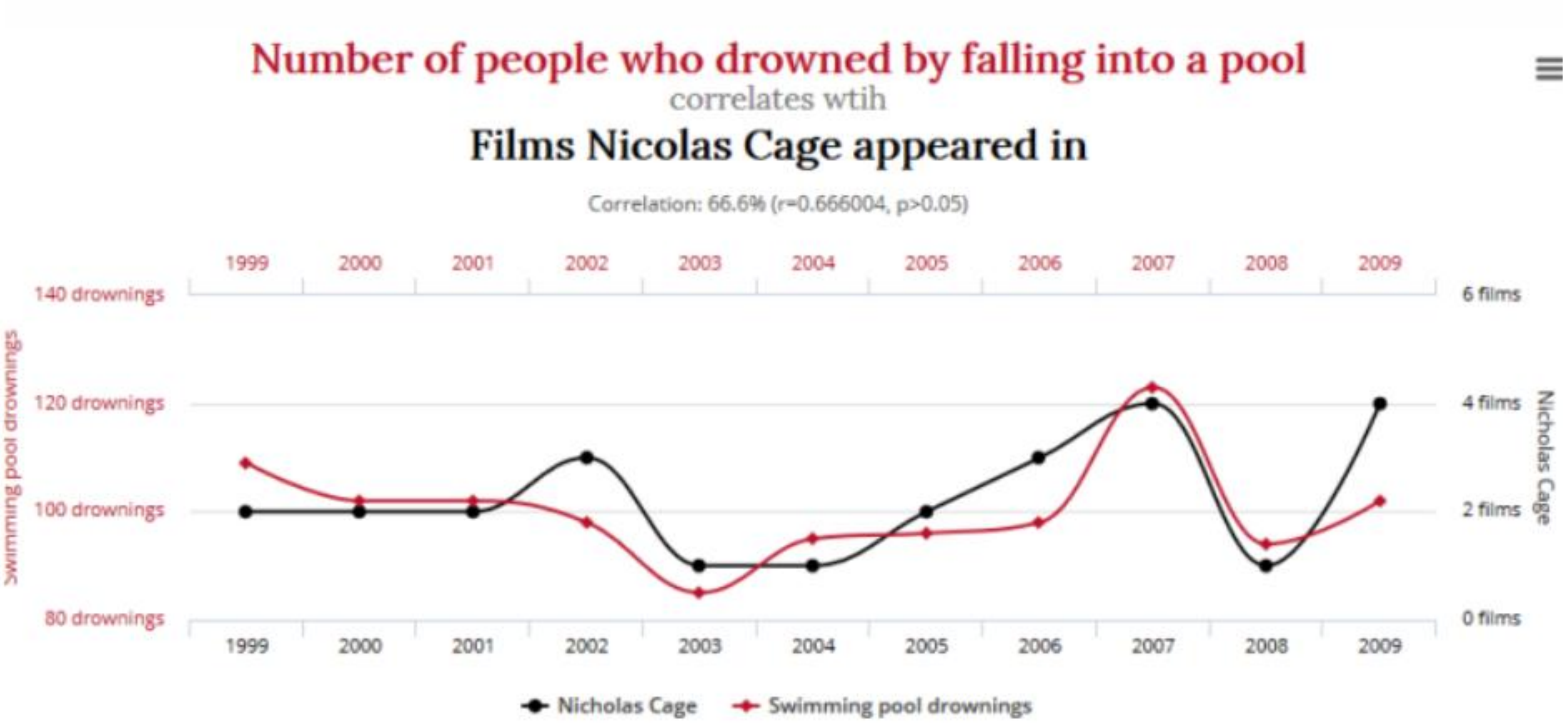
Pearson R:

$$R(x, y) = \frac{Cov(x,y)}{Std(x)Std(y)}$$

Correlation (vs *causation*)



Correlation (vs *causation*)



Correlation (vs *causation*)