

# Workshop 7

COMP20008 Elements of Data Processing

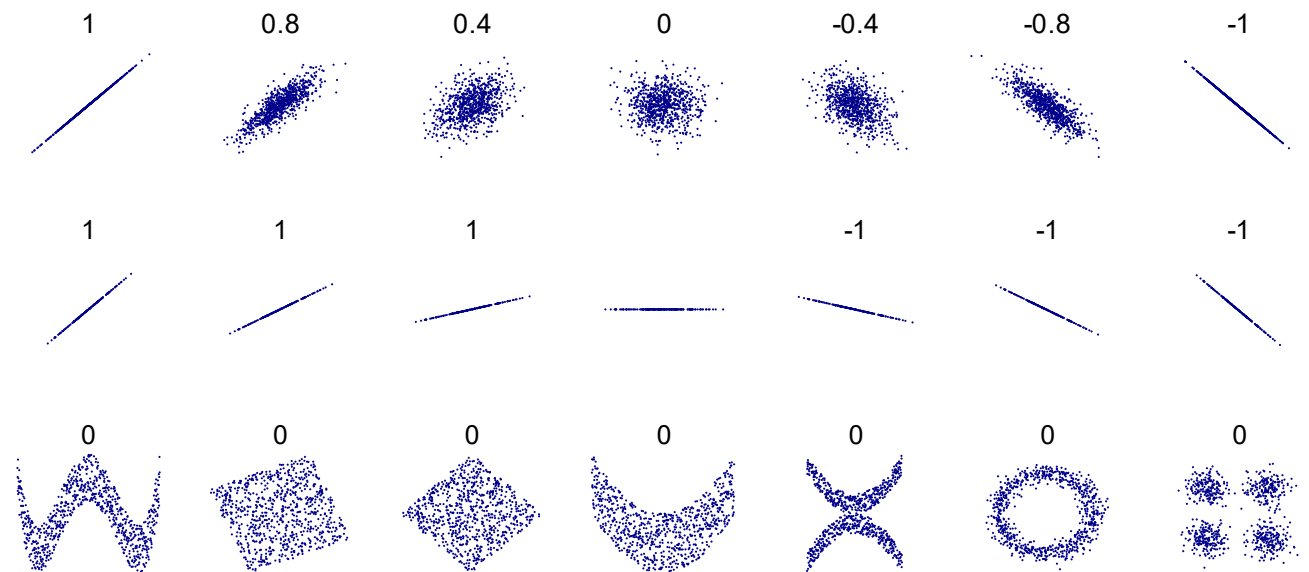
# Learning outcomes

By the end of this class, you should be able to:

- explain the meaning of *correlation* and how it is measured
- compute the *Pearson correlation coefficient* by hand and in Python
- compute *mutual information* by hand and in Python

# Review: What is correlation?

- A statistical relationship between two variables
- Doesn't have to be a *linear* relationship
- Can be measured in different ways, e.g.
  - mutual information
  - Pearson correlation coefficient
- E.g. income and education are correlated



# Pearson correlation coefficient (PCC)

## Statistical definition (not examinable)

The PCC for a pair of random variables  $X$  and  $Y$  is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}}$$

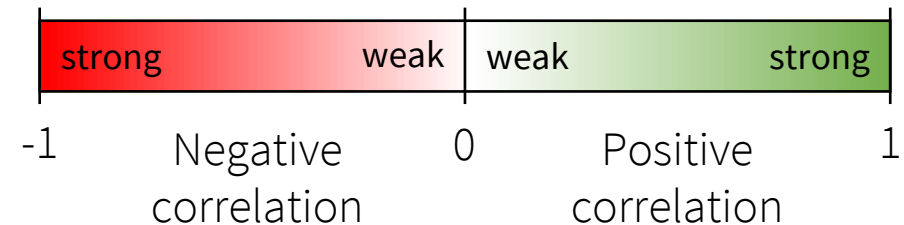
- Measures *linear* correlation
- Many interpretations
  - standardised covariance
  - standardised slope of regression line

## Sample definition

Given observations  $\{(x_i, y_i)\}_{i=1 \dots n}$  the sample PCC is

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

sample mean  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$



# Q1: PCC

Compute the PCC between *average steps per day* (***X***) and *average resting heart rate* (***Y***)

Person ID	Avg. Steps per day ( $x_i$ )	Avg. resting heart rate ( $y_i$ )
1	1000	100
2	2500	105
3	3000	80
4	5000	77
5	6000	74
6	9000	70
7	11000	65
8	14000	63
9	18000	62
10	19000	61
11	19500	60.5
12	22000	55

# Q1: PCC

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Compute the PCC between *average steps per day (X)* and *average resting heart rate (Y)*

Step 1:

Compute the means

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{130000}{12} \approx 10833.3$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{872.5}{12} \approx 72.7083$$

Person ID	Avg. Steps per day ( $x_i$ )	Avg. resting heart rate ( $y_i$ )
1	1000	100
2	2500	105
3	3000	80
4	5000	77
5	6000	74
6	9000	70
7	11000	65
8	14000	63
9	18000	62
10	19000	61
11	19500	60.5
12	22000	55

# Q1: PCC

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Step 2:  
Compute  
the sums

Person ID	Avg. Steps per day ( $x_i$ )	Avg. resting heart rate ( $y_i$ )	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \times (y_i - \bar{y})$
1	1000	100	-9833.33	27.2917	9.66944e7	744.835	-2.68368e5
2	2500	105	-8333.33	32.2917	6.94444e7	1042.75	-2.69097e5
3	3000	80	-7833.33	7.29167	6.13611e7	53.1684	-5.71181e4
4	5000	77	-5833.33	4.29167	3.40278e7	18.4184	-2.50347e4
5	6000	74	-4833.33	1.29167	2.33611e7	1.66840	-6.24306e3
6	9000	70	-1833.33	-2.70833	3.36111e6	7.33507	4.96528e3
7	11000	65	166.667	-7.70833	2.77778e4	59.4184	-1.28472e3
8	14000	63	3166.67	-9.70833	1.00277e7	94.2517	-3.07431e4
9	18000	62	7166.67	-10.7083	5.13611e7	114.668	-7.67431e4
10	19000	61	8166.67	-11.7083	6.66944e7	137.085	-9.56181e4
11	19500	60.5	8666.67	-12.2083	7.51111e7	149.043	-1.05806e5
12	22000	55	11166.7	-17.7083	1.24694e8	313.585	-1.97743e5
Sum	130000	872.5	0	0	6.16167e8	2736.23	-1.12883e6

# Q1: PCC

Step 3:

Substitute intermediate results into formula

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{-1.12883 \times 10^6}{\sqrt{6.16167 \times 10^8} \sqrt{2736.23}} \\ &= -0.8694 \end{aligned}$$



## Q2: Interpretation of PCC

Does a sample PCC of  $-0.8694$  imply doing *more* steps per day will cause one's resting heart rate to *decrease*?

## Q2: Interpretation of PCC

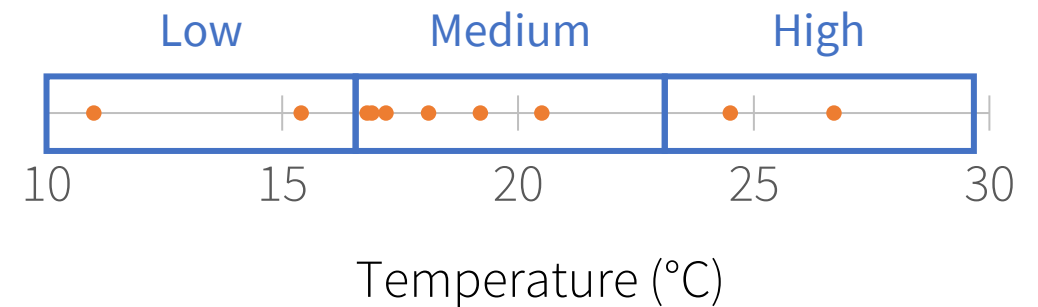
Does a sample PCC of  $-0.8694$  imply doing *more* steps per day will cause one's resting heart rate to *decrease*?

- There's a strong negative correlation, but correlation *does not* imply causation
- There might be a *confounding variable* responsible for the correlation  
e.g. high blood pressure might cause high heart rate and less physical activity
- Also need to be careful about the data. Is there sufficient data? Is it unbiased?

# Discretisation

- Convert continuous variables to discrete variables
- Useful for density estimation (we'll use it to estimate mutual information)
- Various methods:
  - Equal frequency binning
  - Equal width binning
  - Custom method based on domain knowledge

ID	Temp	Temp (Disc)
1	15.4	Low
2	16.9	Medium
3	20.5	Medium
4	24.5	High
5	18.1	Medium
6	17.2	Medium
7	16.8	Medium
8	19.2	Medium
9	11	Low
10	26.7	High




## Q3: Equal-frequency discretisation

Apply 3-bin equal-frequency discretisation to *average steps per day* ( $X$ ) and 4-bin equal-frequency discretisation to *average resting heart rate* ( $Y$ ).


Show the values of the discretised features.

# Q3: Equal-frequency discretisation

Person ID	Avg. Steps per day (X)
1	1000
2	2500
3	3000
4	5000
5	6000
6	9000
7	11000
8	14000
9	18000
10	19000
11	19500
12	22000

  
**Step 1:**  
Sort column


Person ID	Avg. Steps per day (X)
1	1000
2	2500
3	3000
4	5000
5	6000
6	9000
7	11000
8	14000
9	18000
10	19000
11	19500
12	22000

  
**Step 2:**  
Split rows  
into 3 blocks


Person ID	Avg. Steps per day (X)	X (discrete)
1	1000	1
2	2500	1
3	3000	1
4	5000	1
5	6000	2
6	9000	2
7	11000	2
8	14000	2
9	18000	3
10	19000	3
11	19500	3
12	22000	3

# Q3: Equal-frequency discretisation

Person ID	Avg. resting heart rate (Y)
1	100
2	105
3	80
4	77
5	74
6	70
7	65
8	63
9	62
10	61
11	60.5
12	55

  
Step 1:  
Sort column

Person ID	Avg. resting heart rate (Y)
12	55
11	60.5
10	61
9	62
8	63
7	65
6	70
5	74
4	77
3	80
1	100
2	105

  
Step 2:  
Split rows  
into 4 blocks

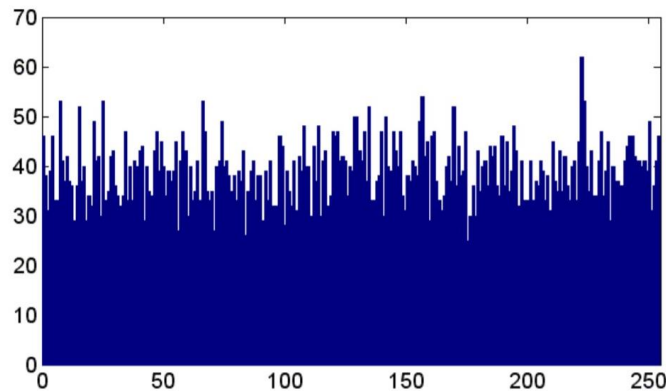
Person ID	Avg. resting heart rate (Y)	Y (discrete)
12	55	1
11	60.5	1
10	61	1
9	62	2
8	63	2
7	65	2
6	70	3
5	74	3
4	77	3
3	80	4
1	100	4
2	105	4

# Q3: Equal-frequency discretisation

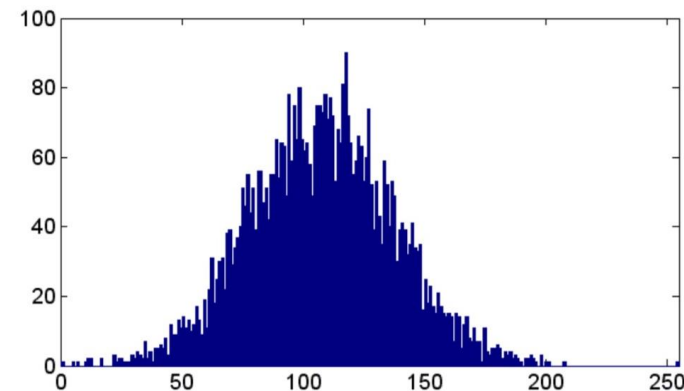
Person ID	Avg. Steps per day (X)	Avg. resting heart rate (Y)	X (discrete)	Y (discrete)
1	1000	100	1	4
2	2500	105	1	4
3	3000	80	1	4
4	5000	77	1	3
5	6000	74	2	3
6	9000	70	2	3
7	11000	65	2	2
8	14000	63	2	2
9	18000	62	3	2
10	19000	61	3	1
11	19500	60.5	3	1
12	22000	55	3	1

# Entropy

- Scalar quantity  $H(X)$  associated with a random variable  $X$
- Interpretation: measures the average level of “information” / “surprise” / “uncertainty” in the outcomes of  $X$



High entropy



Low entropy



# Entropy of a discrete random variable

Entropy:

$$H(X) = - \sum_{x \in \mathcal{X}} p_x \log p_x$$

Conditional entropy:

$$H(Y|X) = \sum_{x \in \mathcal{X}} p_x H(Y|X = x)$$

Interpretation: average  
information given we  
know the outcome of  $X$

where  $p_x$  is the relative frequency of category  $x$

## Q4: Entropy

Compute  $H(X)$ ,  $H(Y)$ ,  $H(Y|X)$  and  $H(X|Y)$  where  $X$  is the discretised avg. steps per day and  $Y$  is the discretised avg. resting heart rate

Person ID	X (discrete)	Y (discrete)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

# Q4: Entropy

Compute  $H(X)$ ,  $H(Y)$ ,  $H(Y|X)$  and  $H(X|Y)$  where  $X$  is the discretised avg. steps per day and  $Y$  is the discretised avg. resting heart rate

$$\begin{aligned} H(X) &= -\sum_{x=1}^3 p_x \log p_x \\ &= -\frac{4}{12} \log \frac{4}{12} - \frac{4}{12} \log \frac{4}{12} - \frac{4}{12} \log \frac{4}{12} \\ &= -3 \left( \frac{4}{12} \log \frac{4}{12} \right) \\ &= 1.585 \end{aligned}$$

$x$	$p_x$
1	4/12
2	4/12
3	4/12

Person ID	X (discrete)	Y (discrete)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

# Q4: Entropy

$$\begin{aligned} H(Y) &= - \sum_{y=1}^4 p_y \log p_y \\ &= - \frac{3}{12} \log \frac{3}{12} - \frac{3}{12} \log \frac{3}{12} - \frac{3}{12} \log \frac{3}{12} - \frac{3}{12} \log \frac{3}{12} \\ &= -4 \left( \frac{3}{12} \log \frac{3}{12} \right) \\ &= 2 \end{aligned}$$

$y$	$p_y$
1	3/12
2	3/12
3	3/12
4	3/12

Person ID	X (discrete)	Y (discrete)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

# Q4: Entropy

$x$	$p_x$
1	4/12
2	4/12
3	4/12

$$\begin{aligned}
 H(Y|X) &= \sum_{x=1}^3 p_x H(Y|X=x) \\
 &= \frac{4}{12} \times 0.811 + \frac{4}{12} \times 1 + \frac{4}{12} \times 0.811 \\
 &= 0.874
 \end{aligned}$$

Person ID	X (discrete)	Y (discrete)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

$$H(Y|X=2) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} = 1$$

$x$	$y$	$p_y$
1	1	0/4
1	2	0/4
1	3	1/4
1	4	3/4

$x$	$y$	$p_y$
2	1	0/4
2	2	2/4
2	3	2/4
2	4	0/4

$x$	$y$	$p_y$
3	1	3/4
3	2	1/4
3	3	0/4
3	4	0/4

$$H(Y|X=1) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0.811$$

$$H(Y|X=3) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0.811$$

# Q4: Entropy

$y$	$p_y$
1	3/12
2	3/12
3	3/12
4	3/12

$$\begin{aligned}
 H(X|Y) &= \sum_{y=1}^4 p_y H(X|Y=y) \\
 &= \frac{3}{12} \times 0 + \frac{3}{12} \times 0.918 + \frac{3}{12} \times 0.918 + \frac{3}{12} \times 0 \\
 &= 2 \left( \frac{3}{12} \times 0.918 \right) \\
 &= 0.459
 \end{aligned}$$

$$H(X|Y=3) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.918$$

$x$	$y$	$p_x$
1	4	3/3
2	4	0/3
3	4	0/3

$$H(X|Y=4) = 0$$

$x$	$y$	$p_x$
1	3	1/3
2	3	2/3
3	3	0/3

$$H(X|Y=2) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.918$$

$x$	$y$	$p_x$
1	2	0/3
2	2	2/3
3	2	1/3

$$H(X|Y=1) = 0$$

$x$	$y$	$p_x$
1	1	0/3
2	1	0/3
3	1	3/3

Person ID	X (discrete)	Y (discrete)
1	1	4
2	1	4
3	1	4
4	1	3
5	2	3
6	2	3
7	2	2
8	2	2
9	3	2
10	3	1
11	3	1
12	3	1

# Mutual information

- Measures *non-linear* correlation
- $\text{MI}(X, Y)$  denotes the mutual information of  $X$  and  $Y$
- Interpretations:
  - the amount of information obtained about  $X$  from observing  $Y$
  - the price paid for encoding  $X, Y$  as independent variables when they're not
- $\text{MI}(X, Y) \geq 0$ 
  - A value of 0 means  $X$  and  $Y$  are independent
  - Larger values indicate stronger dependence



# Mutual information

Mutual information can be expressed in terms of entropy  $H(X)$  and conditional entropy  $H(X|Y)$ :

$$\begin{aligned} \text{MI}(X, Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

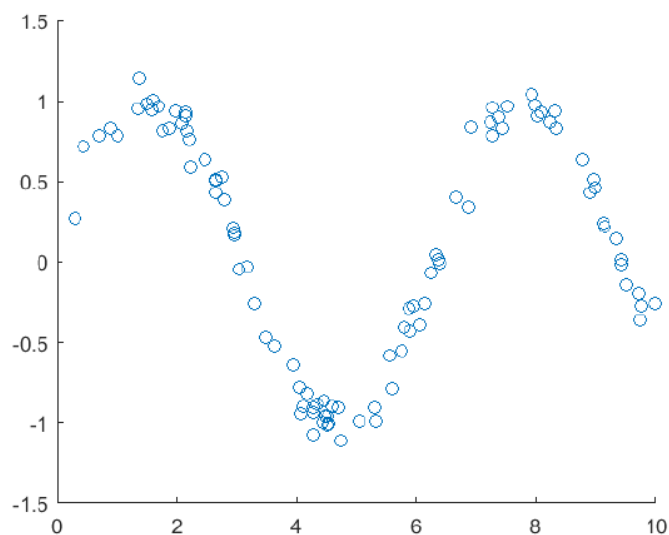
There's also a normalized version:

$$\text{NMI}(X, Y) = \frac{\text{MI}(X, Y)}{\min\{H(X), H(Y)\}}$$

Varies between 0 and 1

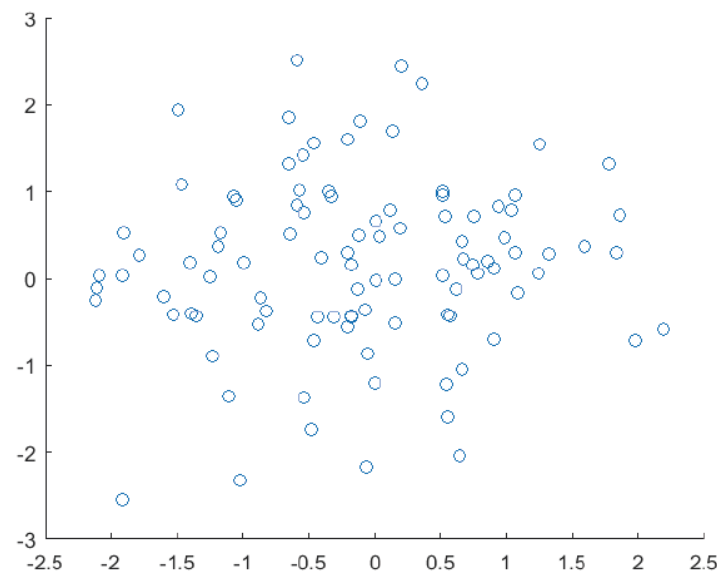


# Mutual information vs. Pearson correlation



PCC: -0.086

NMI: 0.43



PCC: 0.08

NMI: 0.009

## Q5: Mutual information

Compute the mutual information between discretised *average steps per day* ( **$X$** ) and discretised *average resting heart rate* ( **$Y$** )

## Q5: Mutual information

Compute the mutual information between discretised *average steps per day* ( $X$ ) and discretised *average resting heart rate* ( $Y$ )

Substitute earlier results:

$$\text{MI}(X, Y) = H(X) - H(X|Y) = 2 - 0.874 = 1.126$$

$$\text{MI}(X, Y) = H(Y) - H(Y|X) = 1.585 - 0.459 = 1.126$$