COMP229: Introduction to Data Science
Lecture 14: Correlation and scatterplots

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan

- Scatterplot
- Normalisation
- Variance, covariance, correlation
- Correlation vs causation

Reminder: scalar (dot) product

The *scalar (dot)* product of vectors $\vec{u} = (u_1, \ldots, u_n)$ and $\vec{v} = (v_1, \ldots, v_n)$ in $\mathbb{R}^n$ is $\vec{u} \cdot \vec{v} = \sum\limits_{i=1}^{n} u_i v_i = |\vec{u}| \cdot |\vec{v}| \cos \alpha$, where α is the angle between $\vec{u}, \vec{v}$.

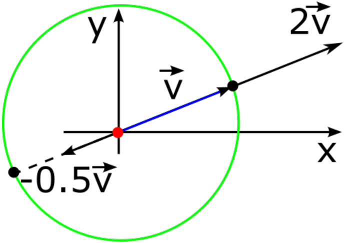
$\vec{u}, \vec{v}$ are orthogonal if and only if $\vec{u} \cdot \vec{v} = 0$.

Cauchy's inequality $|(\vec{u} \cdot \vec{v})| \leqslant |\vec{u}| \cdot |\vec{v}|$ or $\left| \sum\limits_{i=1}^{n} u_i v_i \right|^2 \leqslant \left(\sum\limits_{i=1}^{n} u_i^2 \right) \left(\sum\limits_{i=1}^{n} v_i^2 \right)$.

Equivalence classes of $s\vec{v}$

Let vectors $\vec{u} \sim \vec{v}$ be related if $\vec{u} = s\vec{v}$ for $s \in \mathbb{R}$. This is an equivalence relation for $s \neq 0$.

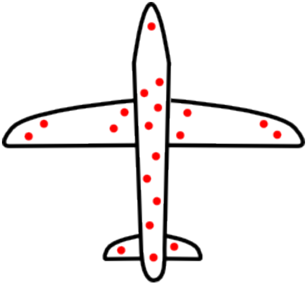
For the equivalence that allows any $s > 0$, any vector $\vec{v} \neq \vec{0}$ has a canonical representative $\frac{\vec{v}}{|\vec{v}|}$ of length 1. The set of equivalence classes is the unit circle S^1 and the class of the zero vector $\{\vec{0}\}$.



If we allow all $s \neq 0$, diametrically opposite points of S^1 should be identified : or identify the endpoints of the angle interval $[0, \pi]$.

How can statistics save lives?

Statistician Abraham Wald plotted locations of all bullet holes in airplanes returning from combat. As data accumulated only few spots on the plane had no bullet holes.
So which areas we should cover?

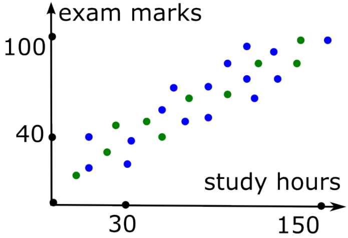


Wald suggested that the armour is added only to these few spots without holes, because other areas with holes were not life-threatening. Visualise data!

The scatterplot of data points

Definition 14.1. If each data object has two quantitative variables (features, descriptors), say x_i, y_i for the i-th object, the **scatterplot** of n data points consists of the n points (x_i, y_i) in the plane.

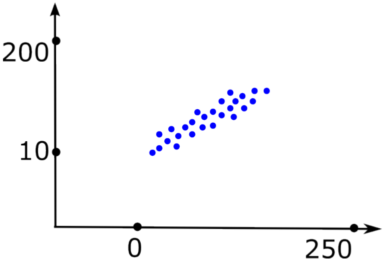
For m variables per object, the scatterplot is in $\mathbb{R}^m$.
What can you conclude from the plot below?

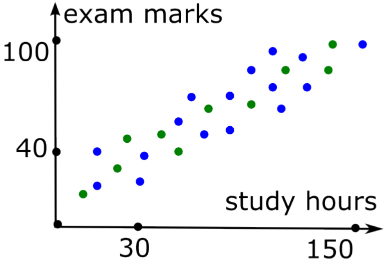


If a variable is categorical (yes/no or only few values), each value can be represented by its own colour.

When a visualisation may not help

The plot on the right below hints that there is a correlation between study hours and exam marks: more study hours often lead to better marks.





Is a correlation stronger in the left plot? To make the analysis rigorous, we define a new concept.

Data normalisation

Real data often comes from different sources and requires a normalisation, e.g. scaling real values to $[0,1]$ or shifting so that the average is 0.

Let measurements be in the range $[p,q]$. How should you map the variable $x \in [p,q]$ to $[0,1]$?

Problem 14.2. Find a function $f(x) = ax + b$ that bijectively (one-to-one) maps $[p,q]$ to $[0,1]$.

You may assume that $f(p) = 0$ and $f(q) = 1$.

1-variable normalisation

Solution 14.2. Shift $[p, q]$ to $[0, q-p]$ and divide by $q-p$ to shorten the range, so $f(x) = \dfrac{x-p}{q-p}$.

The coefficients in $f(x) = ax + b$ are $a = \dfrac{1}{q-p}$, $b = -\dfrac{p}{q-p}$. Check that $f(p) = 0$, $f(q) = 1$.

There are many other normalisations $[p, q] \to [0, 1]$, e.g. the non-linear functions $f_n(x) = \left(\dfrac{x-p}{q-p}\right)^n$ for any integer $n \geqslant 1$. Check that $f_n(p) = 0$, $f_n(q) = 1$.

Var, Covar and Corr

Let variables x, y have n samples x_i, y_i, $i = 1, \ldots, n$ with sample means $\bar{x}$ and $\bar{y}$.

Variance for a random variable X is defined as $Var(X) = \sigma^2 = E[(X - EX)^2] = E(X^2) - (EX)^2$.

Sample variance $Var(x, y) = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Definition 14.3. Covariance for random variables X, Y is $cov(X, Y) = E[(X - EX)(Y - EY)] = E(XY) - (EX)(EY)$.

Sample covariance $q_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

Pearson product-moment correlation $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$.

Sample Pearson correlation $r_{xy} = \frac{q_{xy}}{s_x s_y} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$.

The sample correlation coefficient

In coordinates, $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, the standard deviations are $s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

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

Calculating the sample correlation

Problem 14.4. Find the correlation for these sampled variables

Speed	30	40	50	60	70
Mileage	24	28	31	28	24

Solution 14.4. Average speed $\bar{x} = 50$, mileage $\bar{y} = 27$.

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (-20) \cdot (-3) + (-10) \cdot 1 +$$
$$+ 0 \cdot 4 + 10 \cdot 1 + 20 \cdot (-3) = 0, \text{ hence } r_{xy} = 0.$$

In this example there was no need to compute the deviations:

$$s_x = \sqrt{\frac{1}{4}(2 \cdot 20^2 + 2 \cdot 10^2)} = 10\sqrt{2.5} \text{ and}$$
$$s_y = \sqrt{\frac{1}{4}(2 \cdot (-3)^2 + 2 \cdot 1^2 + 4^2)} = 3.$$

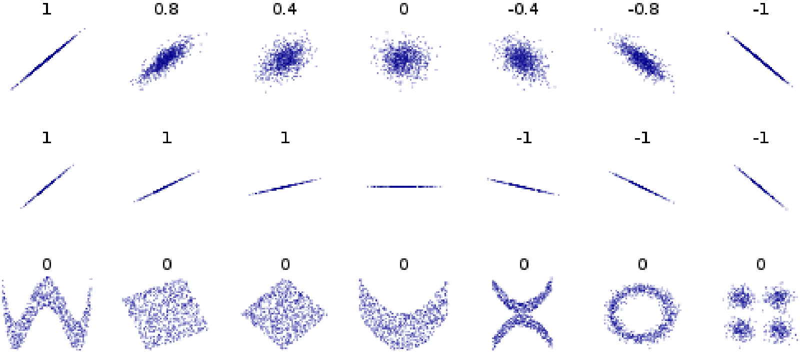
The sign of the correlation

A positive correlation $r_{xy} > 0$ "generally means" that larger values of x correspond to larger values of y, i.e. the scatterplot of (x_i, y_i) is close to a line $y = ax + b$ with a gradient (slope) $a > 0$.

A negative correlation $r_{xy} < 0$ "generally means" that larger values of x correspond to smaller values of y, i.e. the scatterplot of (x_i, y_i) is close to a line $y = ax + b$ with a gradient $a < 0$.

If x, y are independent, then $r_{xy} \approx 0$.
The converse isn't true: x, y can be non-linearly dependent and have $r_{xy} = 0$.

Scatterplots and correlations

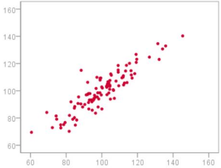


Notice that r_{xy} changes gradually in row 1,
not gradually in row 2 (suddenly from 1 to -1)
and r_{xy} indicates no correlation in row 3 above.

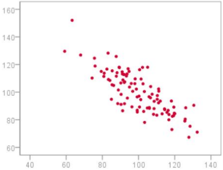
Match scatterplots and correlation

$r_{xy} = -0.8,\ r_{xy} = 0.5,\ r_{xy} = -0.2,\ r_{xy} = 0.9$

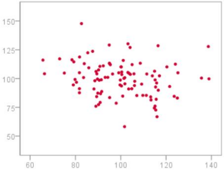
1.



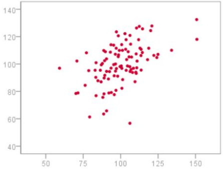
2.



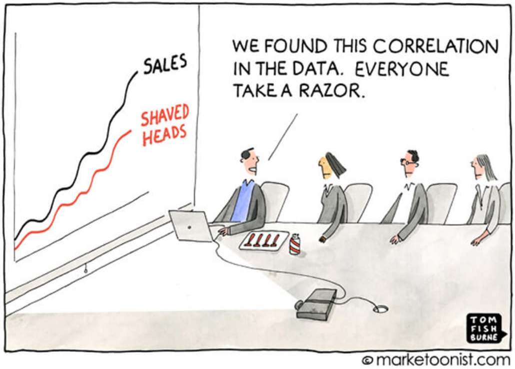
3.



4.

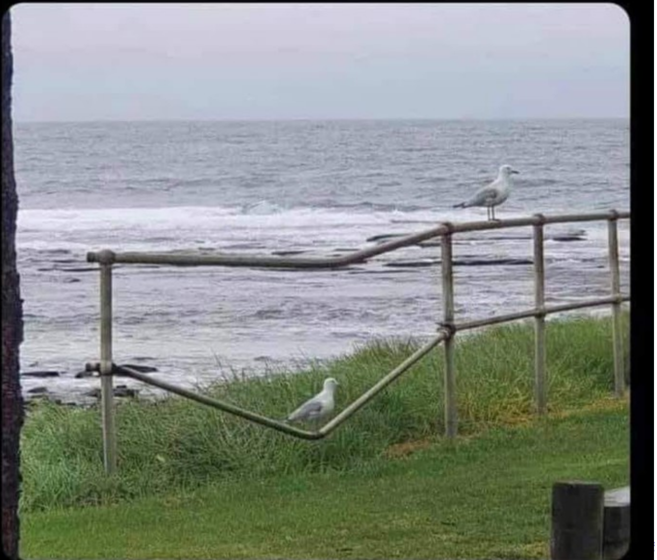


If correlation does exist, be careful!


SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE
© marketoonist.com

For more real life examples, see if chocolate consumption could enhance cognitive function and this.

Correlation Vs. Causation



Linear transformations

Definition 14.5. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called a **linear transformation** if f respects addition and multiplication by scalars, i.e. $f(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v})$ and $f(s\vec{v}) = sf(\vec{v})$ for any $\vec{u}, \vec{v} \in \mathbb{R}^n$, $s \in \mathbb{R}$.

$s = 0$ yields $f(\vec{0}) = \vec{0}$, i.e. any linear f preserves $\vec{0}$.

For $n = m = 1$, $f(x \cdot 1) = x \cdot f(1)$, i.e. $f(x) = ax$ for the constant $a = f(1)$, $x \in \mathbb{R}$. But historically $f(x) = ax + b$ is also called a linear transformation.

For $n, m > 1$, a linear function plus a constant vector $\vec{b}$ is often called **affine**, e.g. $f(\vec{u}) = \vec{u} + \vec{b}$.

The correlation transformation

Claim 14.6. Under any linear transformation $x \mapsto ax + b$ with $a \neq 0$, the correlation r_{xy} is multiplied by

$$\mathrm{sign}(a) = \begin{cases} +1 & \text{for } a > 0 \\ -1 & \text{for } a < 0 \end{cases}.$$

Proof. If x_i are replaced by $ax_i + b$, then

$$r_{ax+b,y} = \frac{\sum_{i=1}^{n}(ax_i - a\bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(ax_i - a\bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{a}{\sqrt{a^2}} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

$\frac{a}{\sqrt{a^2}} r_{xy}$, where $\frac{a}{\sqrt{a^2}} = \mathrm{sign}(a)$ defined above.

Hence r_{xy} has no units of measurement.

The correlation r_{xy} is within $[-1,1]$

Claim 14.7. The sample correlation is symmetric $r_{xy} = r_{yx}$ and is always within the range $[-1,1]$.

Proof. $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$ implies the symmetry.

r_{xy} is invariant under translations $x \mapsto x + b$ as each $x_i - \bar{x}$ remains the same, so we may assume that $\bar{x} = 0 = \bar{y}$. Then $|r_{xy}| \leqslant 1$ is equivalent to

$$|\sum_{i=1}^{n} x_i y_i| \leqslant (n-1)s_x s_y = \sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}.$$

For the vectors $\vec{u} = (x_1, \ldots, x_n)$, $\vec{v} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, it is Cauchy's inequality is $|\vec{u} \cdot \vec{v}| \leqslant |\vec{u}| \cdot |\vec{v}|$, which follows from $\vec{u} \cdot \vec{v} = |\vec{u}| \cdot |\vec{v}| \cdot \cos\alpha$ and $|\cos\alpha| \leqslant 1$. Here α is the angle between $\vec{u}, \vec{v}$.

A linear dependence

Theorem 14.8. If data samples x_i, y_i are **linearly dependent**, i.e. $y_i = ax_i + b$, then $r_{xy} = \mathrm{sign}(a)$.

Proof. The linear transformation $x_i \mapsto ax_i + b = y_i$ multiplies r_{xy} by $\mathrm{sign}(a)$.

After we get two identical samples, the correlation becomes

$$r_{yy} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = 1. \text{ Hence}$$

$r_{xy} = 1/\mathrm{sign}(a) = \mathrm{sign}(a)$.

Is the opposite implication true? We'll see on the next lecture.

Other correlation types

Two other common types of correlation are **rank correlation coefficients: Spearman's and Kendall's τ.**

They measure correlation between *ranks*, hence measure a different type of association.

Problem 14.5. Find correlations for

x	0	10	101	102
y	1	100	500	2000

Solution 14.5. Ranks perfectly match, so rank correlations are 1.

$\bar{x} = 53.25, \bar{y} = 650.25.$ $s_x \approx 56, s_y \approx 925,$

$q_{xy} \approx 38999, r_{xy} \approx 0.75.$

Time to revise and ask questions

- The *scatterplot* consists of points (x_i, y_i) whose coordinates are values of a data object.
- The *sample correlation* between variables x, y is $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \in [-1, 1]$.
- $r_{xy} > 0$ often means that y increases with x.
- $r_{xy} < 0$ often means that y decreases with x.
- $r_{xy} = 0$ means no linear relation between x, y.

Problem 14.6. Find the sample correlation for

x	1	2	3	4	5
y	1	3	5	7	9

Final solution

Solution 14.6. $r_{xy} = 1$ since $y = 2x - 1$. Check!