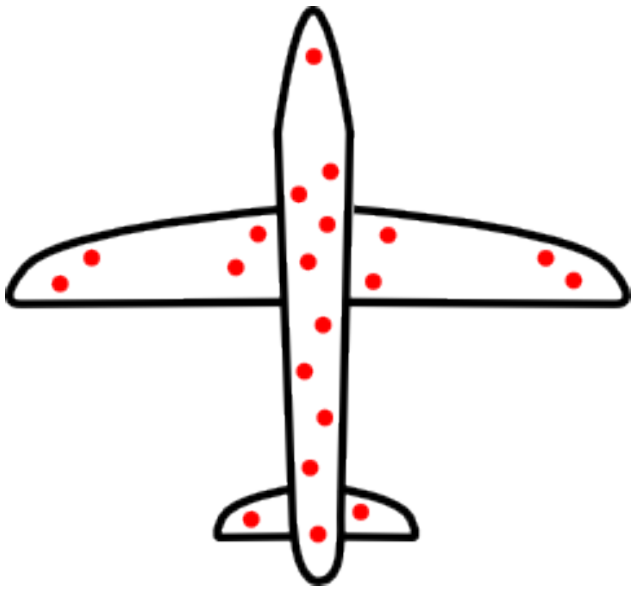
COMP229: Introduction to Data Science
Lecture 14: scatterplots and correlation

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

How can statistics save lives?

Statistician Abraham Wald plotted the location of bullet holes in airplanes returning from combat.

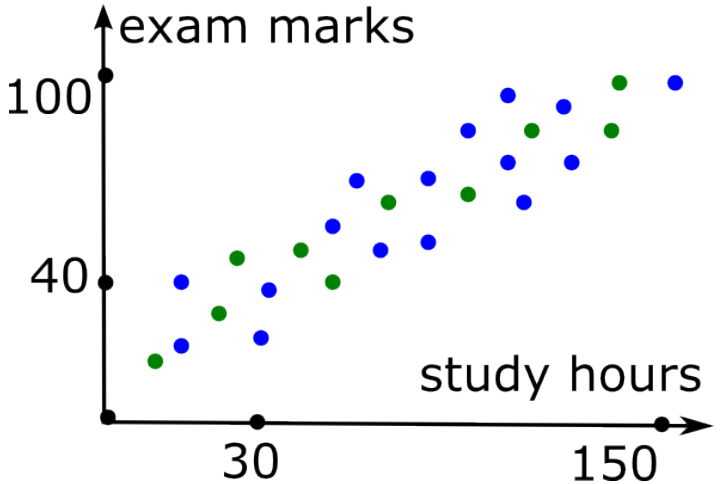


As data accumulated only few spots on the plane had no bullet holes. Then Wald suggested that the armour is added only to these few spots without holes, because other areas with holes were not life-threatening. Visualise data!

The scatterplot of variables

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.

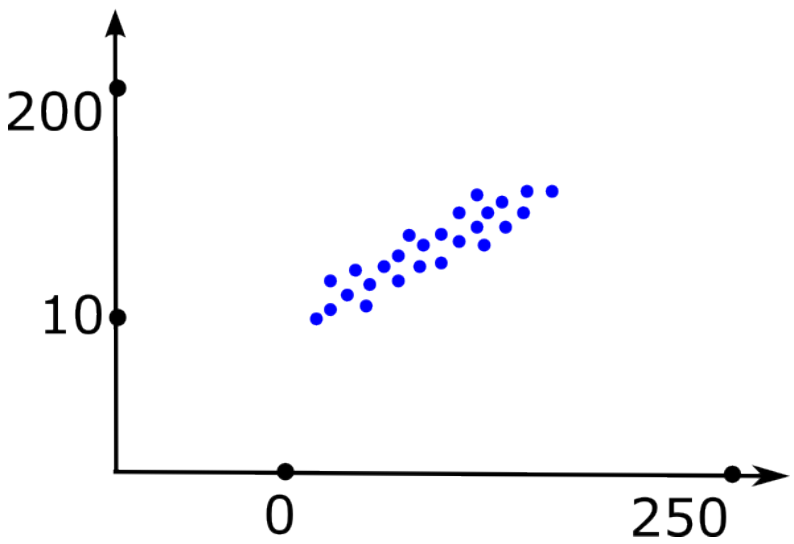
For m variables per object, the scatterplot is in $\mathbb{R}^m$. What can you conclude from the plot below?

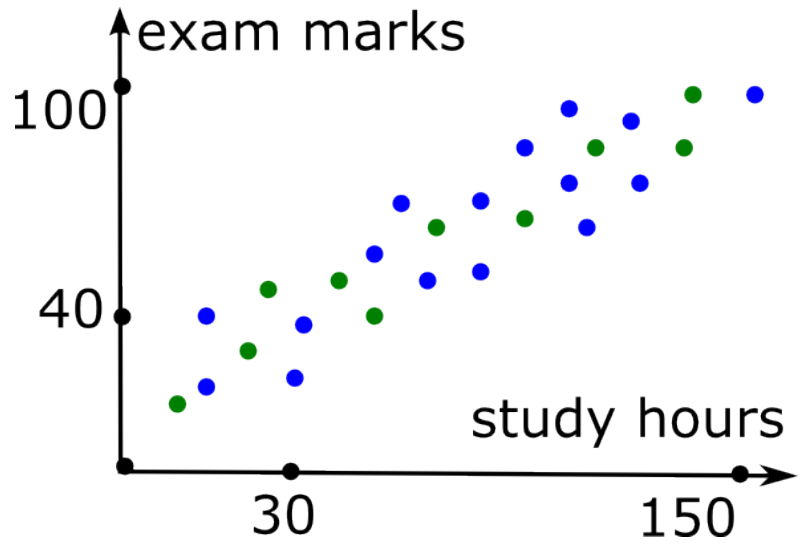


If a variable is categorical (yes/no or only few values), each value can be represented by its own colour.

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 introduce the definition.

The sample correlation coefficient

Definition 14.2. Let 2 variables x, y have n samples x_i, y_i, $i = 1, \ldots, n$. The *sample correlation* between x, y is $r = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$.

Here $\bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$, $\bar{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i$ are sample means.

$s_x = \sqrt{\dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}$, $s_y = \sqrt{\dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (y_i - \bar{y})^2}$ are sample standard deviations. Find the correlation for

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

The sample correlation can be zero

The speed sample $30, 40, 50, 60, 70$ has $\bar{x} = 50$.

The mileage sample $24, 28, 31, 28, 24$ has $\bar{y} = 27$.

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

There was no need to compute the deviations:

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

$$s_y = \sqrt{\frac{1}{4}(2 \cdot (-3)^2 + 2 \cdot 1^2 + 4^2)} = 3.$$

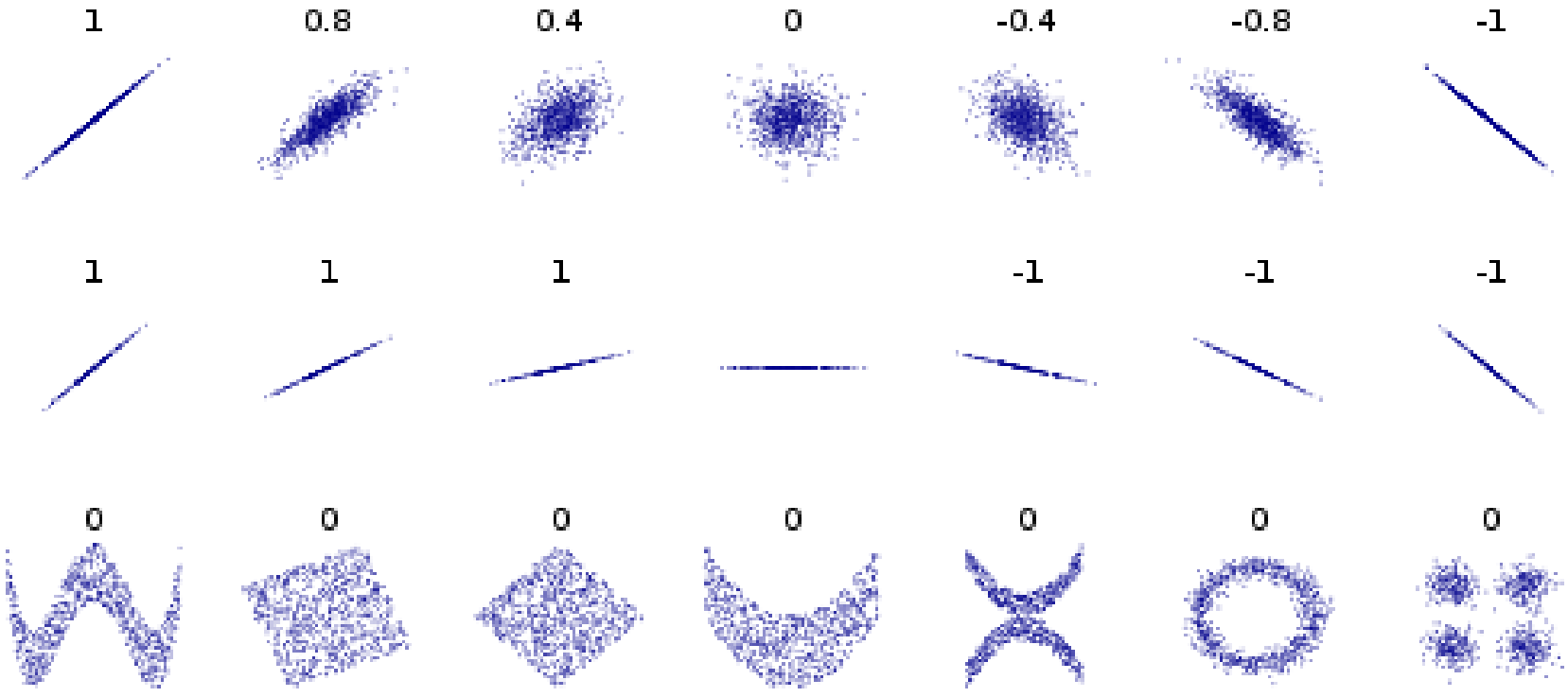
The sign of the correlation

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 $a < 0$.

A relation between a, r_{xy}, s_x, s_y will be later.

Scatterplots and correlations



Notice that r_{xy} changes gradually in row 1, not gradually in row 2 (suddenly from 1 to 0 to -1) and r_{xy} indicates no correlation in row 3 above.

Properties of the sample correlation

Claim 14.3. The sample correlation is symmetric: $r_{xy} = r_{yx}$. The sample correlation is between ± 1.

The sample correlation is invariant under changes of units, hence has no units of measurement.

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. If all x_i are multiplied by a factor t, then $\bar{x}$ and $s_x = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ are multiplied by t, hence r_{xy} remains the same, similarly for y_i.

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

r_{xy} is invariant under translations (shifts), because each $x_i - \bar{x}$ remains the same, so we may assume that $\bar{x} = 0 = \bar{y}$. Then $|r_{xy}| \leq 1$ is equivalent to

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

For the vectors $\vec{u} = (x_1, \ldots, x_n)$, $\vec{v} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$, the above inequality is $|\vec{u} \cdot \vec{v}| \leq |\vec{u}| \cdot |\vec{v}|$,

which follows from $\vec{u} \cdot \vec{v} = |\vec{u}| \cdot |\vec{v}| \cdot \cos\alpha$ and $|\cos\alpha| \leq 1$. Here α is the angle between $\vec{u}, \vec{v}$. $\square$

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Find the sample correlation for

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

Answer to the quiz and summary

Answer. $r_{xy} = 1$, because $y = 2x - 1$. Check!

- 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 $r_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \in [-1, 1]$.
- $r_{xy} > 0$ means that y increases with x.
- $r_{xy} < 0$ means that y decreases with x.
- $r_{xy} = 0$ means no correlation between x, y.