2. Covariance and Correlation

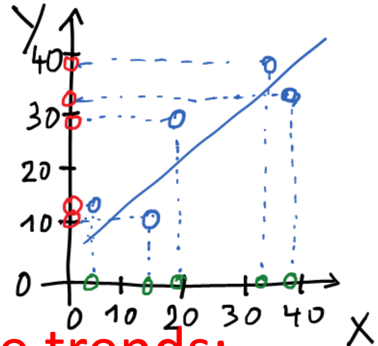
- ✓ Covariance
- ✓ Correlation

Covariance

<u>Covariance can classify three relationships between variables:</u>

1. Relationships with positive trends:

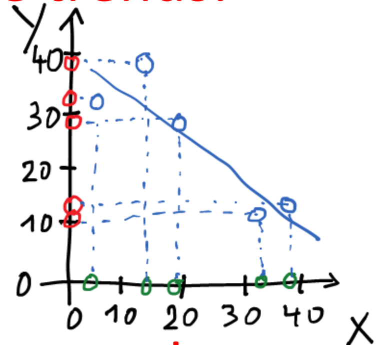
Covariance is POSITIVE



<u>Example:</u>
mRNA of gene X in 5 different cells (X-axis)
*and mRNA of gene Y in **SAME 5 different cells** (Y-axis)*

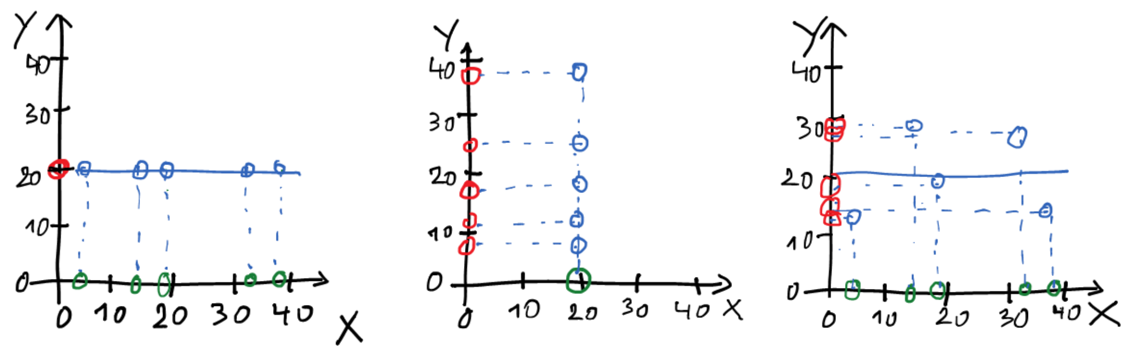
2. Relationships with negative trends:

Covariance is NEGATIVE



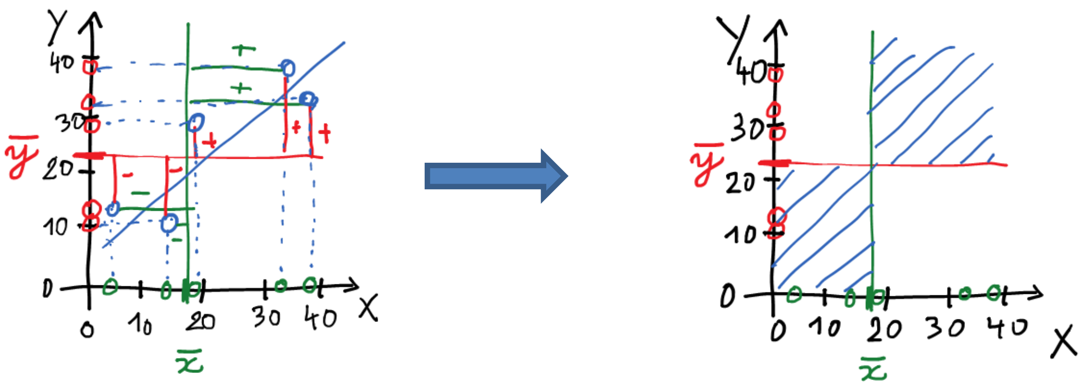
3. No relationship because no trend:

Covariance = 0

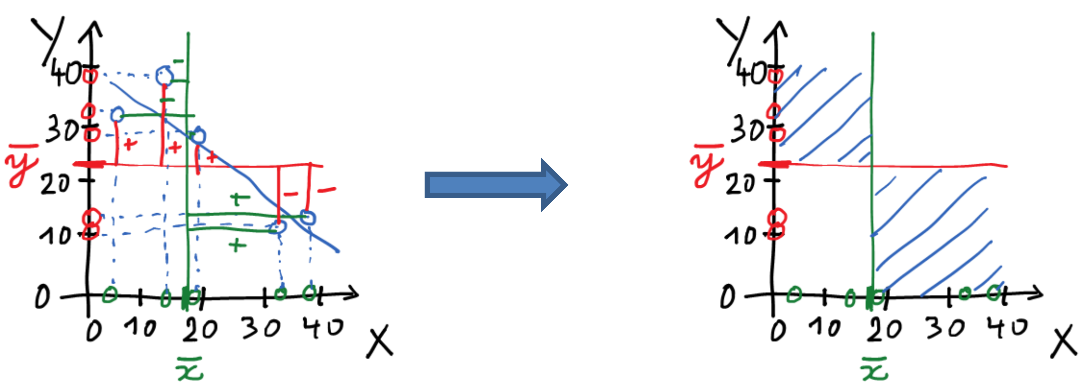


$$Covariance = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{SP_{xy}}{n - 1}$$

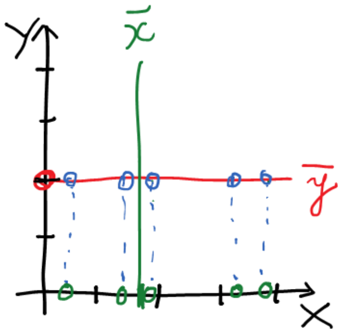
SPxy *refers to the* ***sum of cross products***

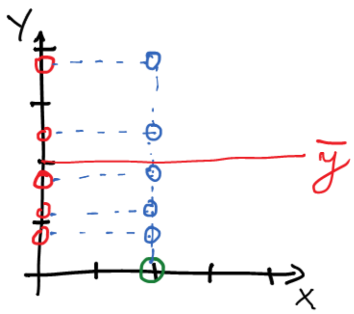


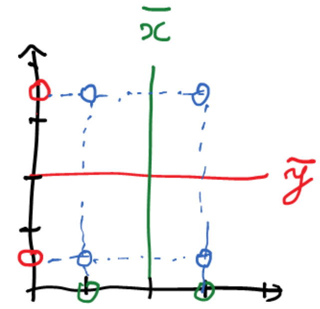
Data in these two quadrants contribute positive values to the total covariance.



Data in these two quadrants contribute negative values to the total covariance.







$(y - \bar{y}) = 0$

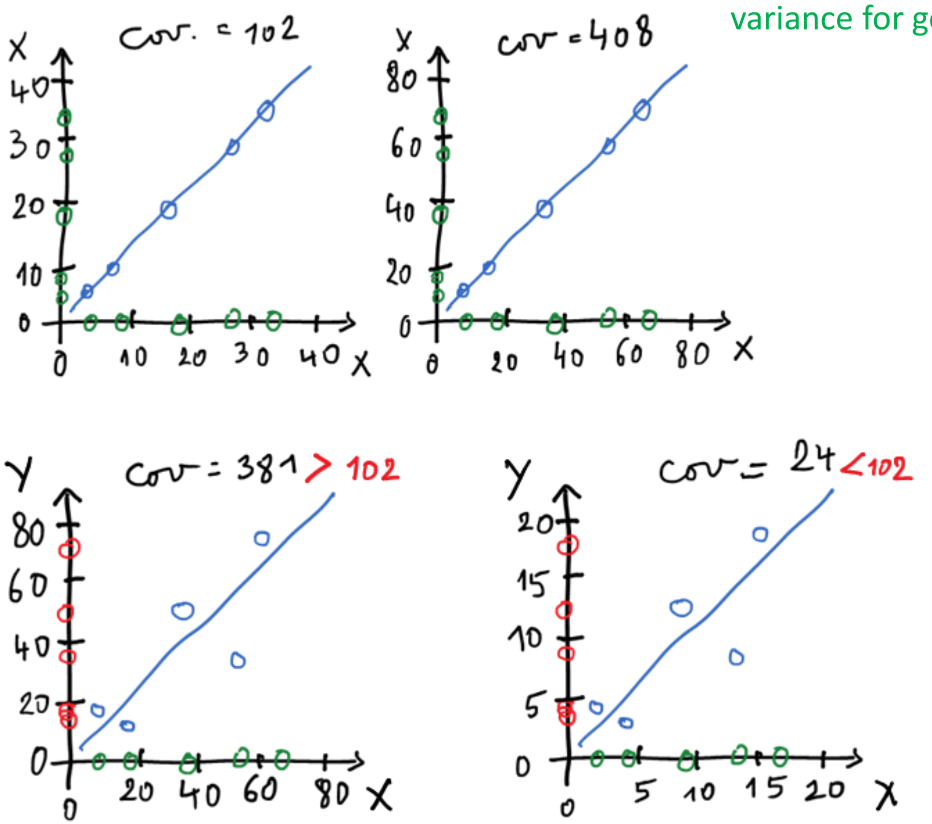
$(x - \bar{x}) = 0$

Positive values canceled out by negative values

NO RELATIONSHIP between X and Y

$$Covariance = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1} = 0$$

Covariance cannot quantify the strength of a relationship



The covariance for gene X with itself is the same thing as the estimated variance for gene X

$$\frac{\sum(x-\bar{x})(y-\bar{x})}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1} = \sigma^2$$

- If you change the scale that the data is on (all values x2), the relationship does not change, BUT the covariance changes!
- Sensitivity to the scale makes that the covariance cannot tell if the data are close to the line representing the relationship, or far from that line.
- Covariance doesn't tell anything about the slope of the line representing the relationship.

COVARIANCE JUST TELLS IF THE RELATIONSHIP IS POSITIVE OR NEGATIVE

→ *Covariance is a computational stepping stone to something that is interesting, like correlation or PCA.*

Correlation

Correlation

- Correlation describes relationships between variables and is NOT sensitive to the scale of the data

→THEREFORE: correlation can quantify the strength of a relationship.

- $\text{Correlation} = \dfrac{\text{Covariance (X,Y)}}{\sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}}$

Numerator: any value between $-\infty$ and $+\infty$, depending on

1) whether the slope of the line that represents the relationship is positive or negative
2) how far the data are spread around the means
3) the scale of the data

$$\text{Correlation} = \frac{\text{Covariance (X,Y)}}{\sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}}$$

"How much of the variance in the data (X and Y) is accounted for by the covariance?"

Denominator: squeezes the covariance to be a number between -1 and $+1$

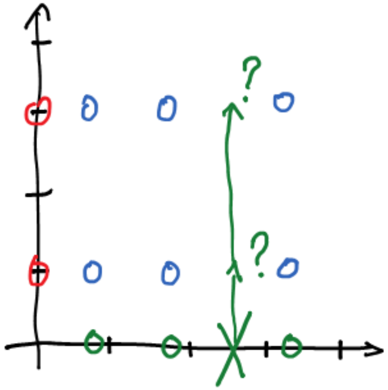
→ensures that the scale of the data does not affect correlation

→Correlations are easier to interpret than covariance.

1. When there is NO RELATIONSHIP between X and Y, none of the variance in the data is accounted for by the covariance (covariance is 0), the correlation is 0.

$$\text{Covariance} = 0$$

$$\Leftrightarrow \text{Correlation} = \frac{0}{\sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}} = 0$$



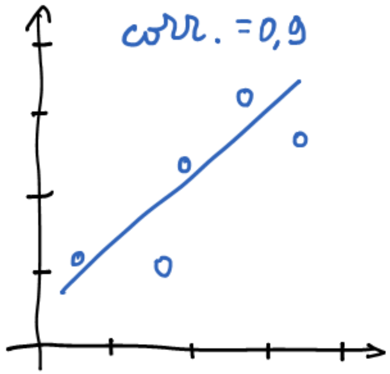
- If correlation = 0

⇔ a value on the X-axis doesn't tell us anything about what to expect on the Y-axis, because there is no reason to chose one value or another.

2. When we observe a trend in the data, BUT the data do not all fall on a straight line, the covariance accounts for some but not all of the variance in the data, and the correlation gets closer to 1 or -1.

$$|\text{Covariance}| < \sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}$$

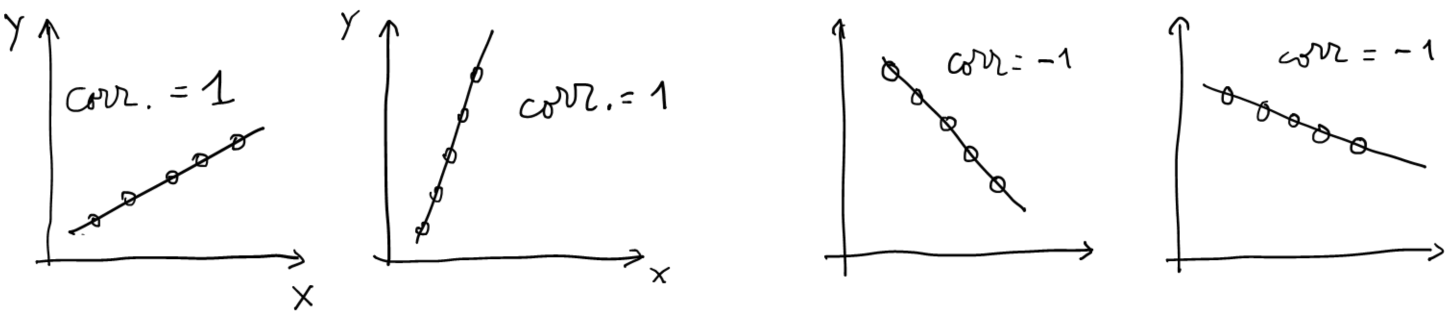
$$\Leftrightarrow -1 < \left[\text{Correlation} = \frac{\text{Covariance (X,Y)}}{\sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}}\right] < 1$$



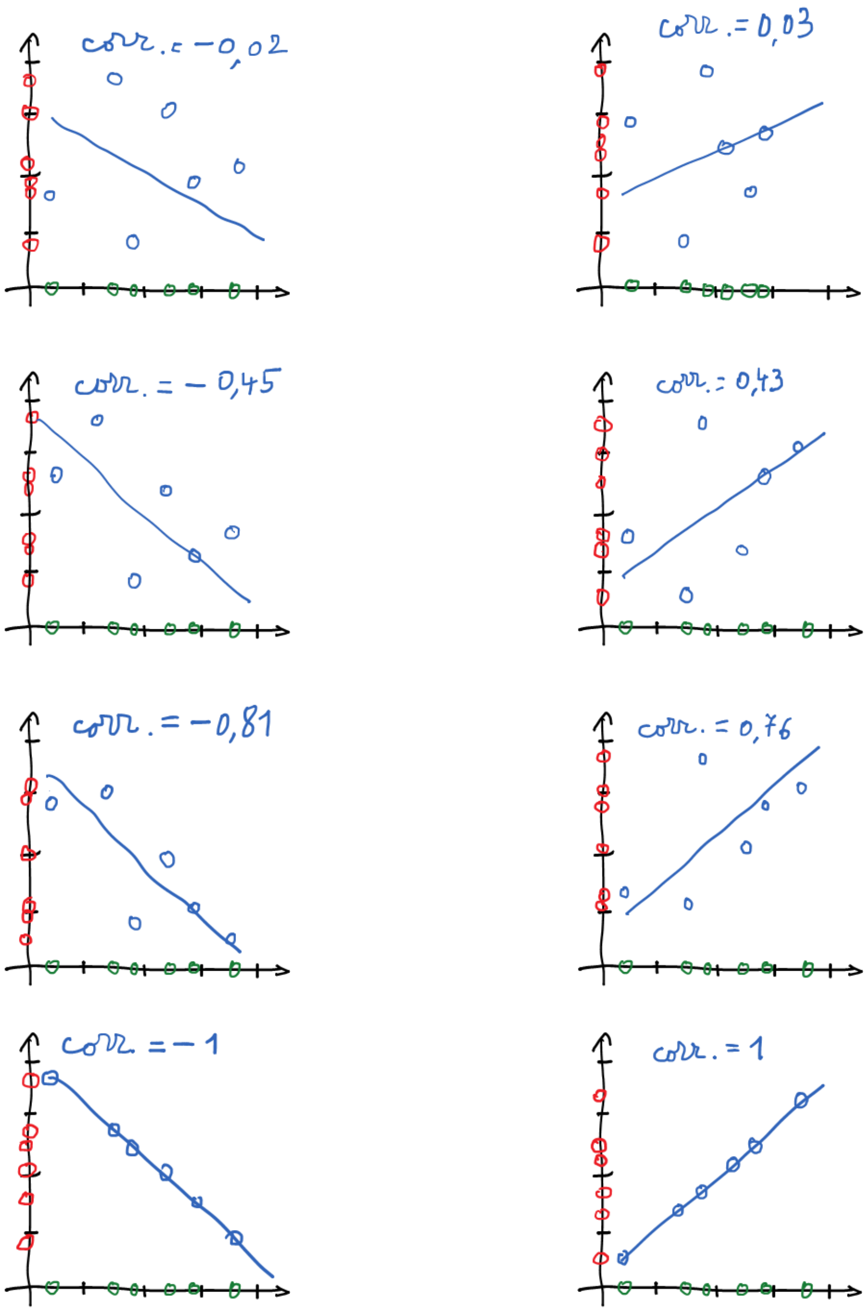
3. When all the data fall on a straight line with positive or negative slope, the covariance accounts for all the variance in the data, and the correlation is 1 or -1.

$$|\text{Covariance}| = \sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}$$

$$\Leftrightarrow \text{Correlation} = \frac{\text{Covariance (X,Y)}}{\sqrt{\text{Variance(X)}}\sqrt{\text{Variance(Y)}}} = \pm 1$$



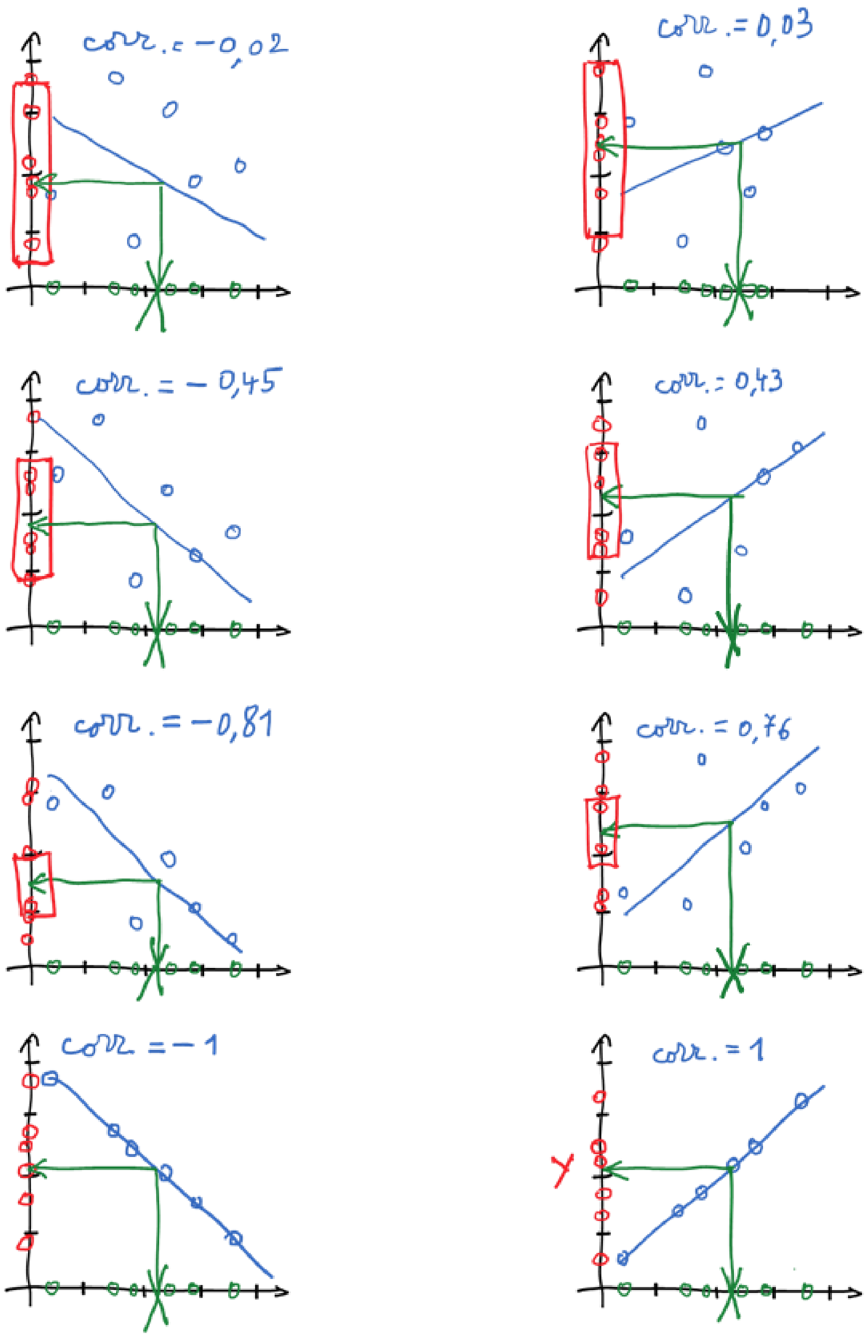
Slope can be large or small



Correlation can quantify the strength of a relationship.

Data closer to the trend line with negative slope
⇔ correlation closer to -1

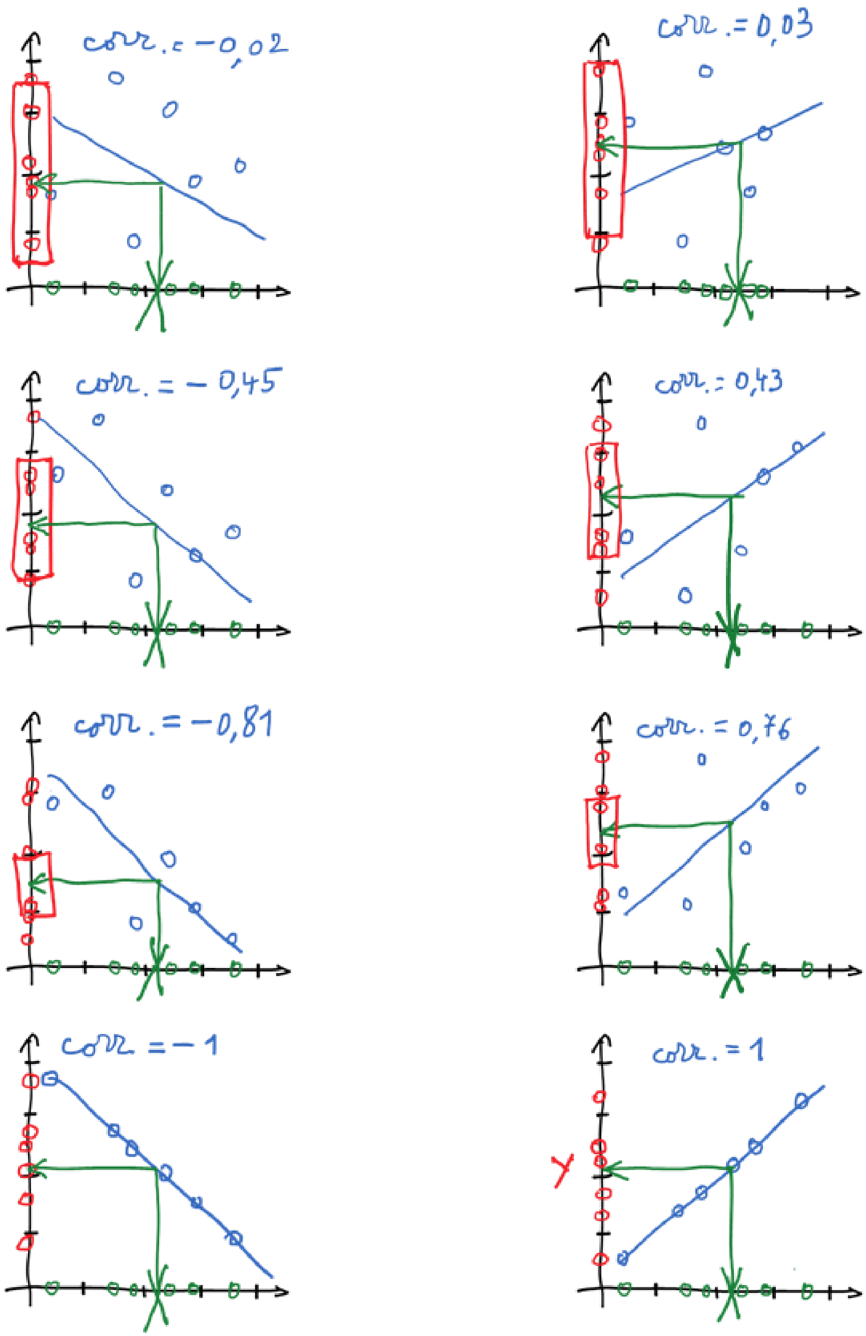
Data closer to the trend line with positive slope
⇔ correlation closer to 1



- If correlation ≠ 0

⇔ the trend can be used to make predictions (educated guesses).

- When correlation values get closer to -1 or 1; the guesses become more refined.

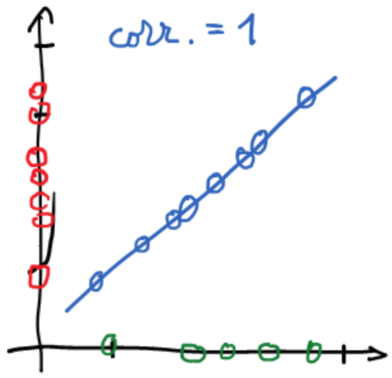


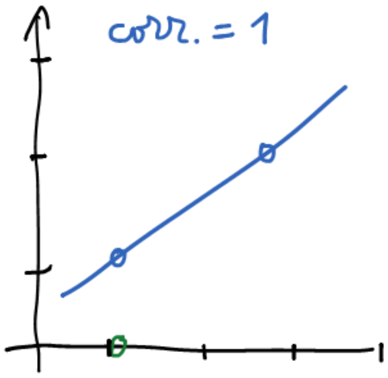
- If we have a new measurement for gene X, then we can use the trend line to predict a range of values for gene Y.

- If we have a new measurement for gene Y, then we can use the trend line to predict a range of values for gene X.

- If the data are closer to the trend line, then
    - →there is a stronger relationship between X and Y.
    - →values for gene X tell us more about gene Y (and *vice versa*).
    - →given a value for gene X, we predict that gene Y falls in a narrower range.

The confidence in how useful the relationship is depends on the amount of data





- When all data are on a straight line, correlation is 1 or -1, regardless of how much data we have.
- **Two datapoints** are ALWAYS on a straight line!

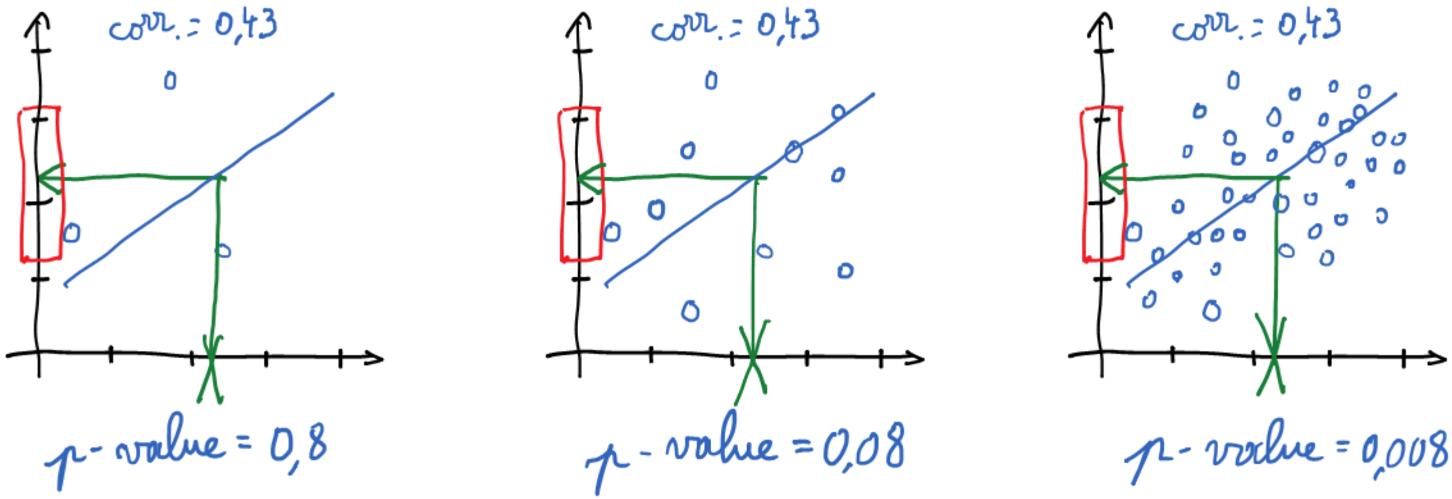
  → correlation is 1 or -1

  → This makes the relationship appear strong

  → But we should not have any confidence in predictions made with this line.
- The probability that we can draw a straight line through a number of points gets smaller with each additional point.

  → The more data we have, the more confidence we have in the predictions we make with the line.

- The **p-value for correlation** tells the probability that randomly drawn dots will result in a similarly strong relationship, or stronger.
- Smaller p-value ⇔ more confidence in the predictions made with the line



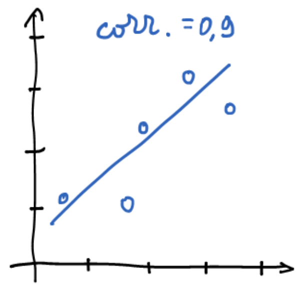
Correlation value 0.43 is small: all three graphs represent BAD GUESSES.
BUT most confidence in the bad guess that came from most data.

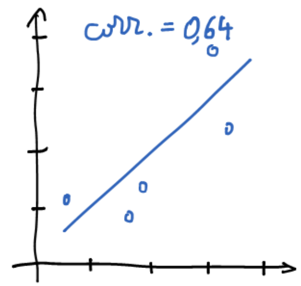
Conclusion:

Even if you have a lot of data and therefore a lot of confidence in your guess (low p-value), if the correlation value is close to 0, your guess will still be bad!

Correlations are not the most easy to interpret.

- Not obvious that the relationship with correlation value 0.9 is twice as good at making predictions as the relationship with correlation value 0.64





- R^2 solves this problem

```
> 0.9^2/0.64^2
[1] 1.977539
```

- R^2 can also quantify relationships that are more complicated than straight lines!

(see later!)

Parameters for populations
(→using Greek letters!)

Covariance between x and y: σ_{xy}

$$\sigma_{xy} = \left(\sum_{i=1}^{N} (y_i - \mu_y)(x_i - \mu_x) \right) \Big/ N$$

can be negative!

Correlation (Pearson's) between two variables, *y and x: ρ*

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^{\,2} \times \sigma_y^{\,2}}}$$

Standardize the covariance

Ranges from -1 to +1; with strong negative correlations near to -1 and strong positive correlations near to +1.

Statistics from the sample

(= estimates of the population parameters!)

Covariance between x and y: *s_{xy}*

$$s_{xy} = \left(\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) \right) \Big/ (n-1)$$

Correlation <u>(Pearson's) </u>between two variables, *y and x: r*

$$r_{xy} = \frac{s_{xy}}{\sqrt{{s_x}^2 \times {s_y}^2}}$$

Ranges from -1 to +1; with strong negative correlations near to -1 and strong positive correlations near to +1.