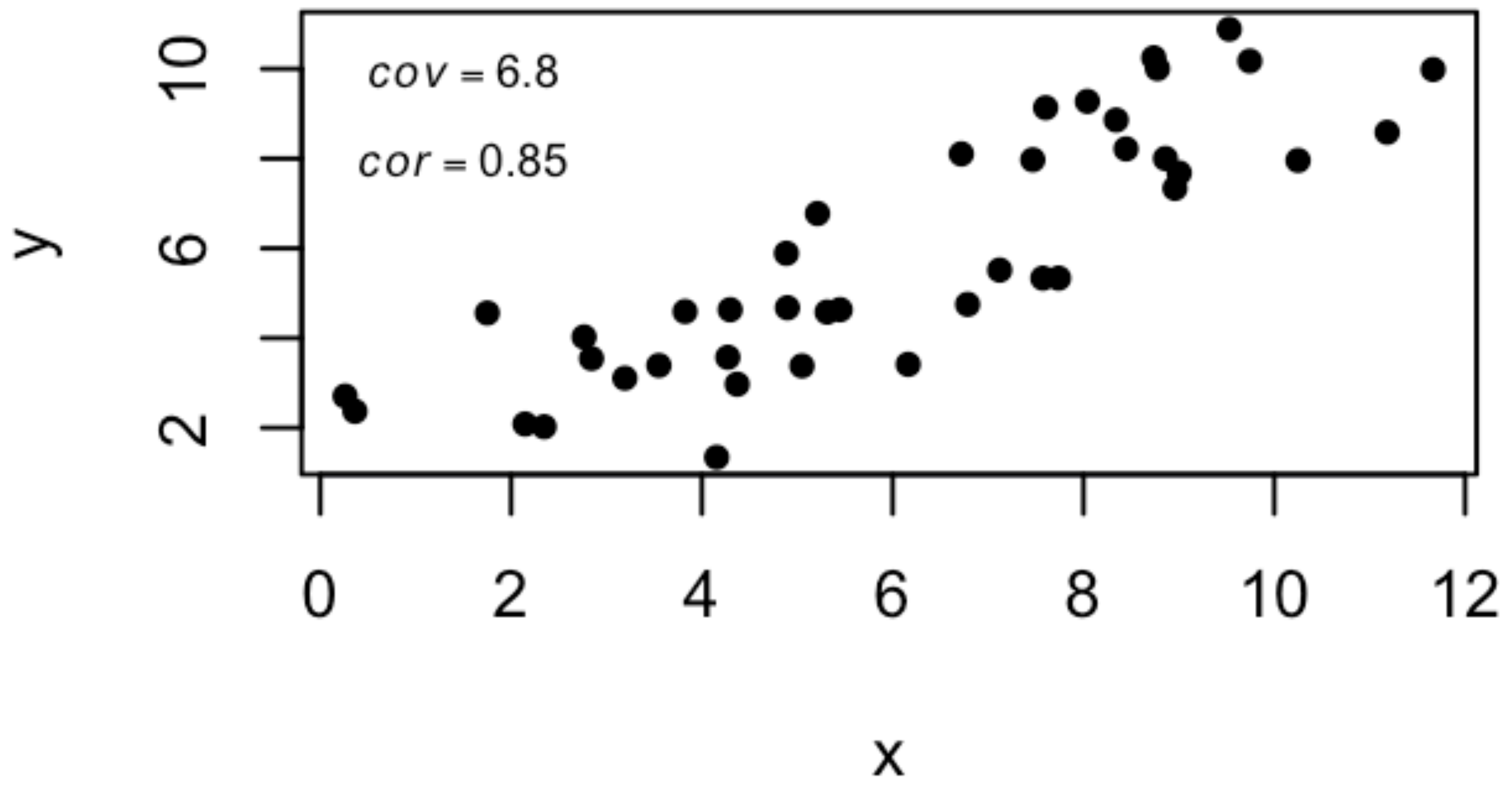
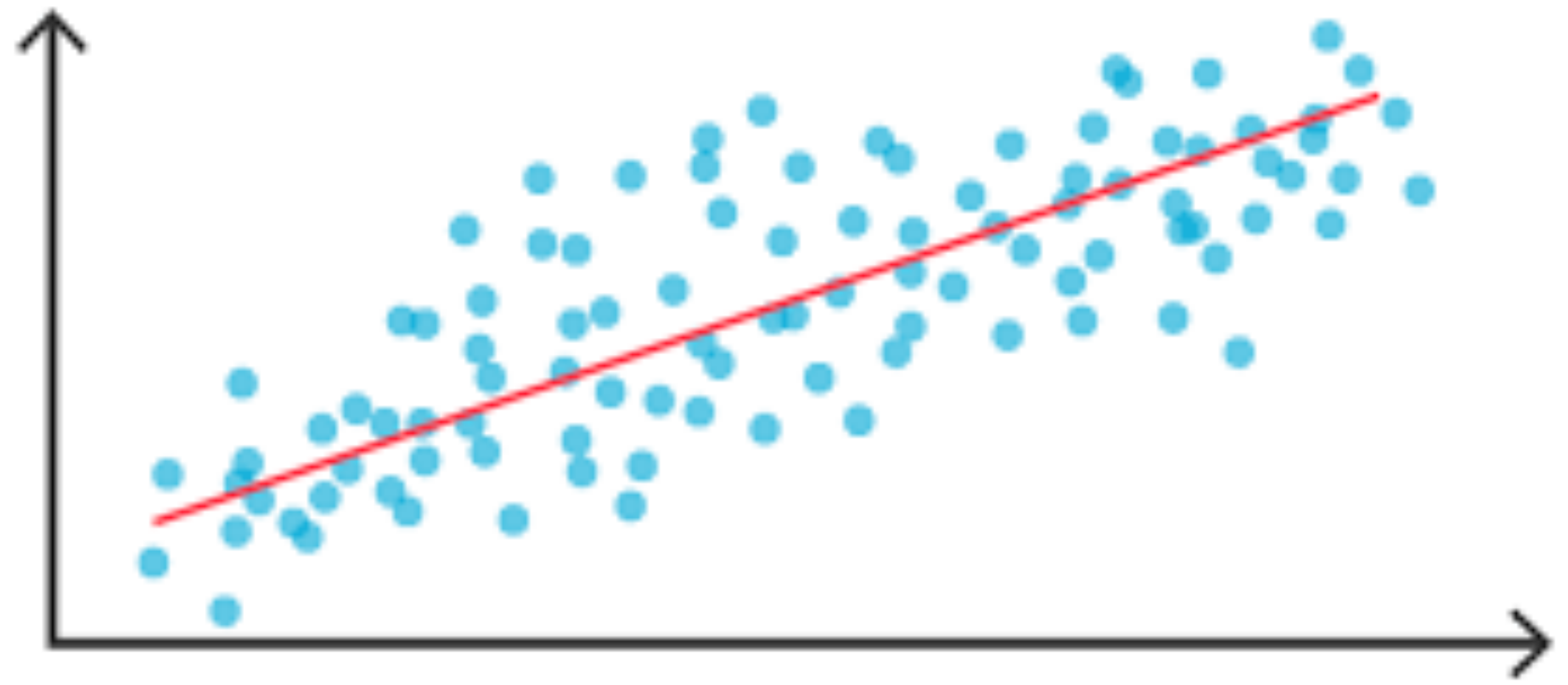
Foundational Statistics

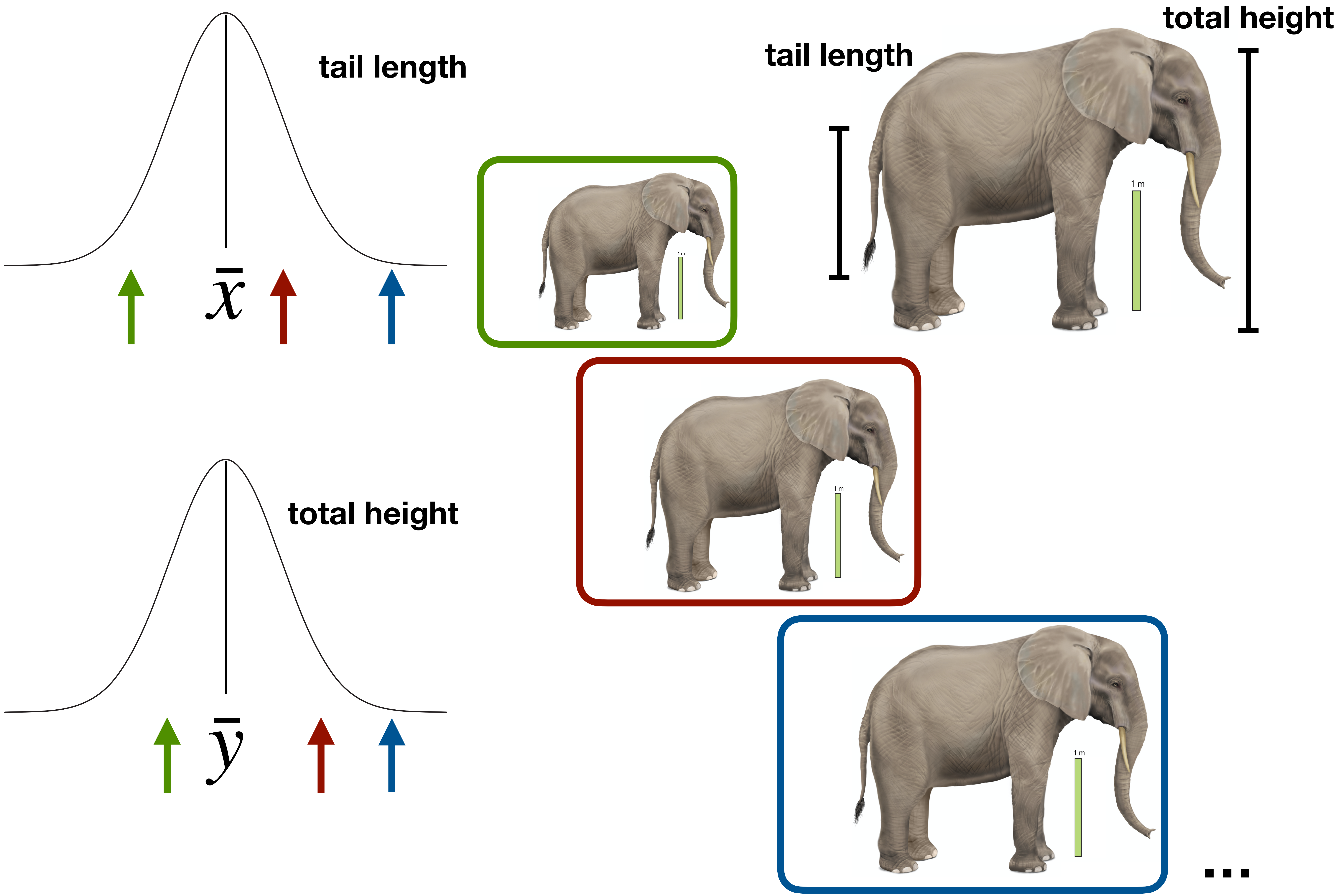
Covariance and Correlation





$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Relationships between numeric variables



Relationships between numeric variables

How do we quantify whether one variable is systematically related to another variable?

The sample covariance:

vars deviate from their means in same dir.: product is positive

vars deviate from their means in opposite dir.: product is negative

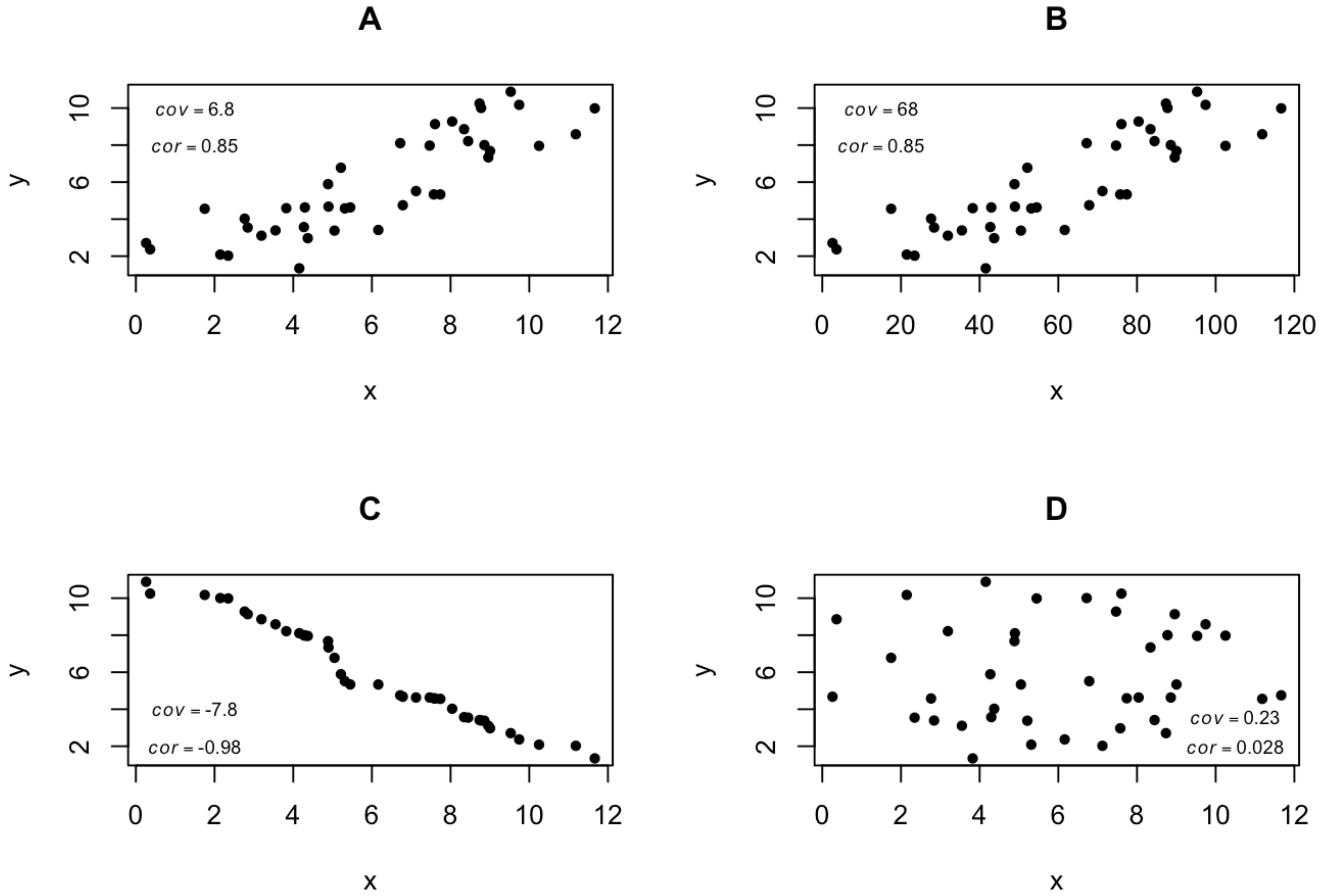
$$cov(x, y) = s_{xy} = \frac{\sum_{i=1}^{n} \overbrace{(x_i - \bar{x})(y_i - \bar{y})}}{n-1}$$

The sample correlation coefficient:

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

The denominator scales *r* so that it ranges from -1 to +1

Covariance and correlation: examples



Hypothesis tests for correlation

$$H_0 : \rho_1 = 0$$

$$H_A : \rho_1 \neq 0$$

One test statistic for this hypothesis test:

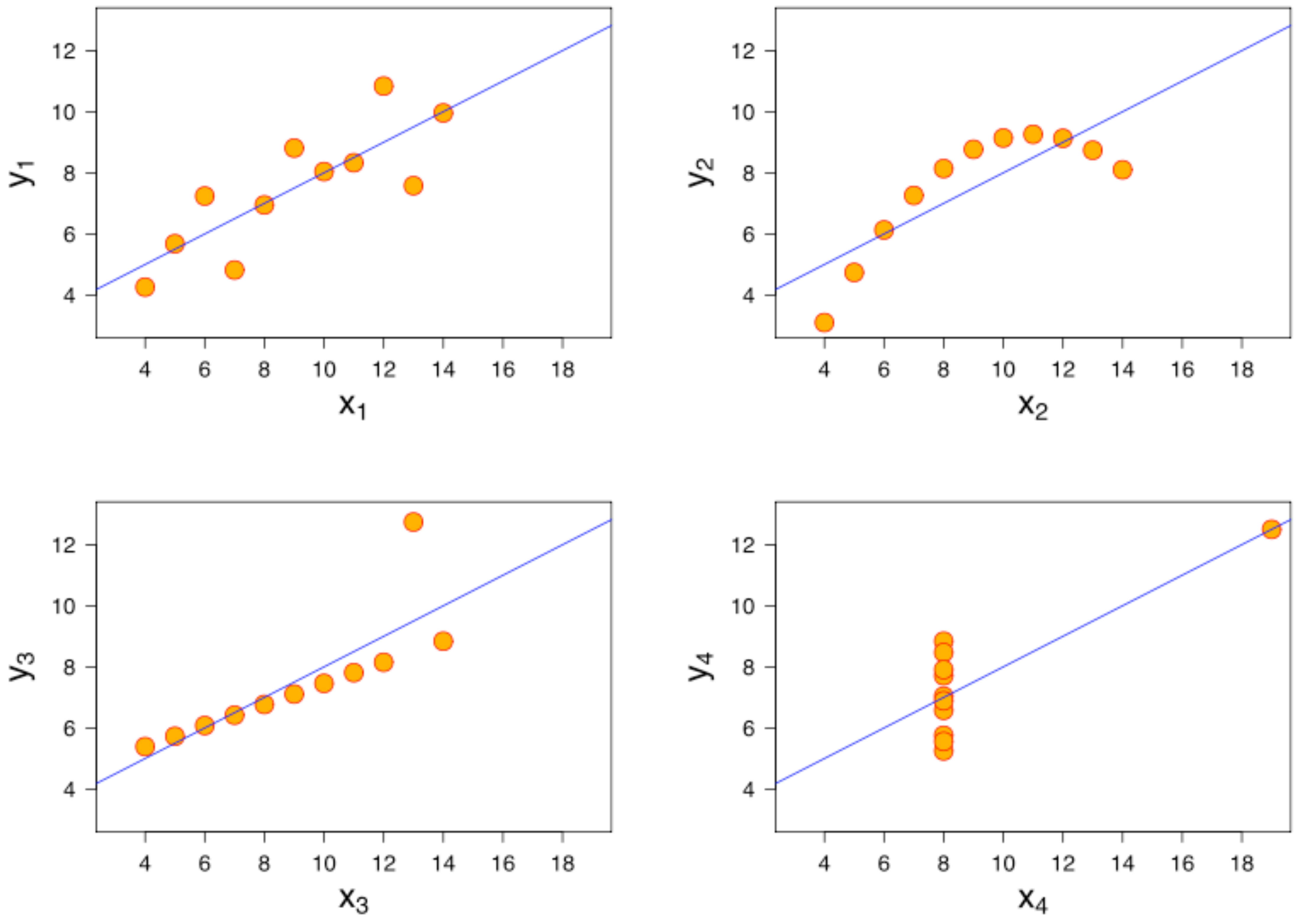
$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Compare to a *t*-distribution with *n* - 2 df
cor.test() function in R will carry out

Assumptions of the test:

1. Relationship mostly linear (no strong curvilinearity)
2. "Bivariate normal": both variables normally dist.

Anscombe’s Quartet: Always plot your data to assess assumptions and guide interpretation!



Mean of *x* in each case
9 (exact)

Variance of *x* in each case
11 (exact)

Mean of *y* in each case
7.50 (to 2 decimal places)

Variance of *y* in each case
4.122 or 4.127 (to 3 decimal places)

Correlation between *x* and *y* in each case
0.816 (to 3 decimal places)

Hypothesis tests for correlation: Nonparametric alternatives

1. **Spearman's rank:** Rank-based, for $n < 30$
2. **Kendall's tau:** Rank-based, for larger sample sizes
3. **Randomization or resampling test**

Key properties of correlation analysis

1. **Indicates directionality**
2. **Strength of relationship is scaled by the variances**
3. **Does not say anything about causation**
4. **Does not say anything about the *steepness* of the relationship (need regression for that!)**