STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 2 tutorial: Correlations, linear regression & bootstrapping

Pearson's *r* vs. Spearman's rho

- We can normalise the covariance by the standard deviations to obtain the *correlation coefficient, r,* sometimes called *Pearson's r* to distinguish it from other types of correlation coefficient :

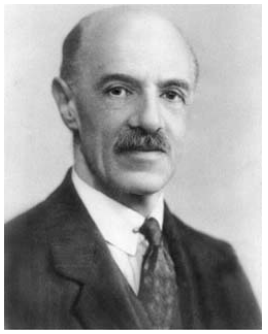
$$r = \frac{s_{xy}}{s_x s_y} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

- Alternatively, rank the x_i and y_i separately in numerical order* and define a difference d_i = rank(x_i) – rank(y_i)

Now we compute *Spearman's ρ* :

$$\rho = 1 - \frac{6\sum_{n=1}^{i} d_i^2}{n(n^2-1)}$$

Karl Pearson
1857-1936



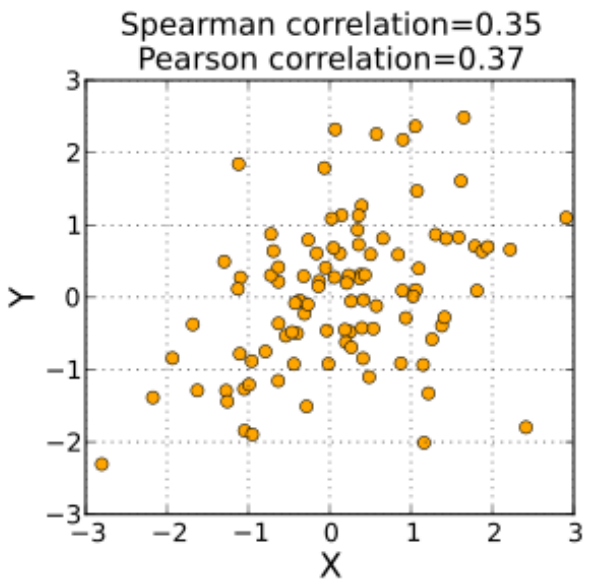
Charles Spearman
1863-1945

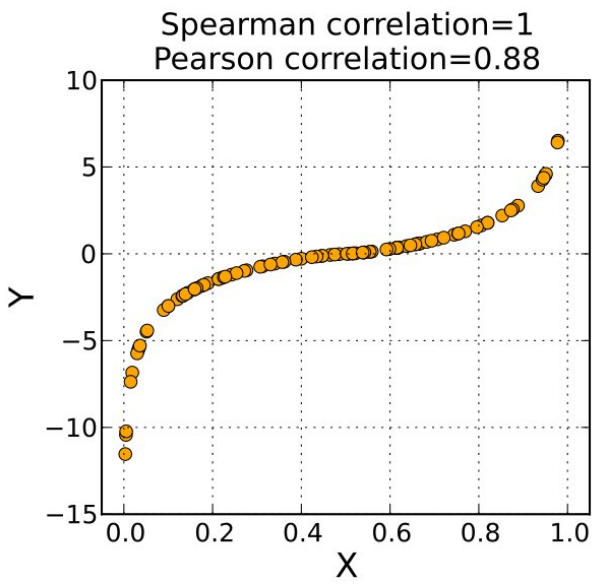
*Note that equal values in the sequence are assigned a rank equal to their average position, e.g. the 4th and 5th highest positions have equal values *x* and are given a rank 4.5 each

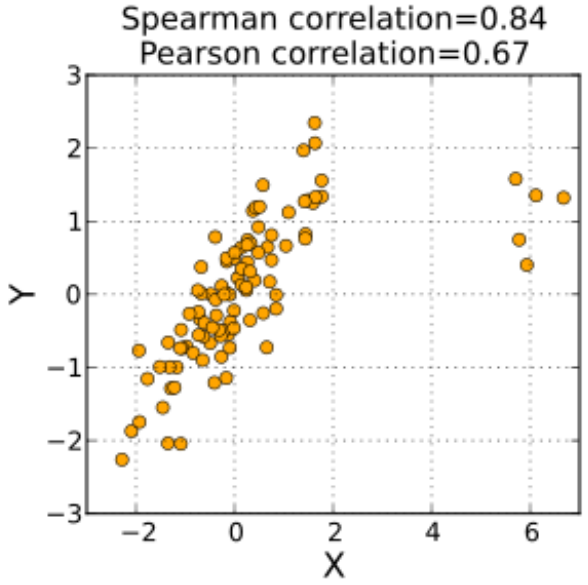
Pearson or Spearman?

- Pearson's *r* is designed to search for *linear correlations* while Spearman's *ρ* is suited to monotonically related variables.

- Spearman's *ρ* is also better able to deal with outliers in the tail of the sample of *x* and *y*, since the contribution of these values to the correlation is limited by their ranks (i.e. irrespective of any large values the outlying data points may have).







Is the correlation real?

- The significance of a correlation, i.e. how unlikely it is that the data are consistent with zero correlation, depends on both the sample size *n* and the size of the correlation co-efficient (*r* or *ρ*).

- Provided that the data are ***independent and identically distributed (i.i.d.)***, the probability of whether a correlation co-efficient is significant can be calculated by transforming to a variable *t* :

$$t = r\sqrt{\frac{n-2}{1-r^2}} \qquad \text{or} \qquad t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

- if the true correlation coefficient is zero, *t* is distributed as a (2-sided) Student's *t*-distribution with *n-2* degrees of freedom.

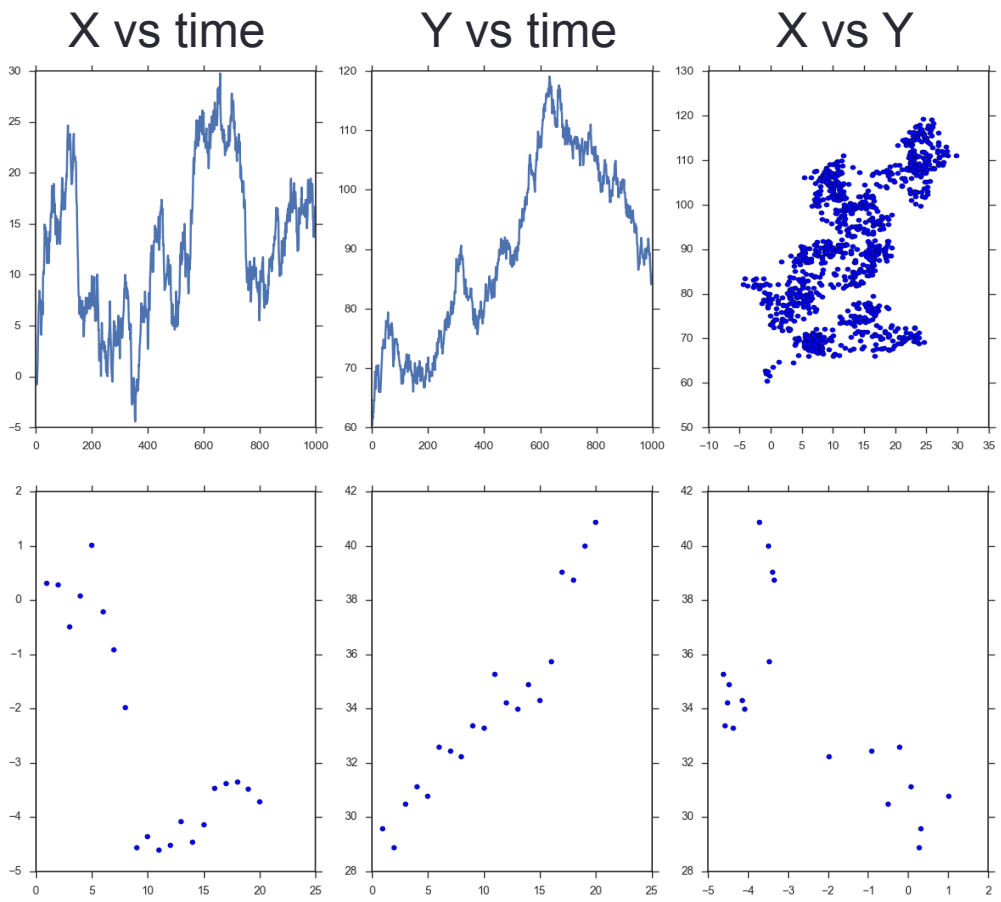
- **However, the measured significance of a correlation does strongly depend on the i.i.d. assumption. We can hardly stress this point enough!!!**

Where things can go really wrong: autocorrelated data

Many measurable quantities vary over time following a random walk – even if two quantities follow completely independent random walks, they can show significant correlations **because the data points in each time-series are not independent of each other** (they are 'autocorrelated')

$$r = 0.539,$$
$$n = 1000,$$
$$P = 2.6 \times 10^{-76}$$

$$r = -0.661,$$
$$n = 20,$$
$$P = 0.0015$$



Simple fits to bivariate data: linear regression

Minimise scatter (*residuals)* around a linear model:

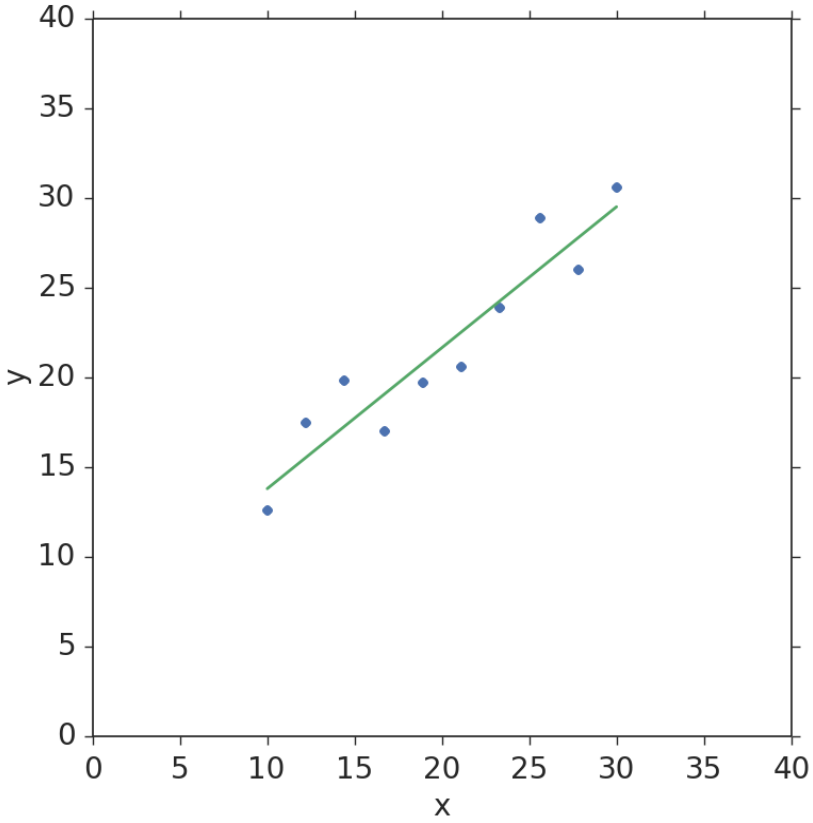
residual = data - model

$$e_i = y_i - (\alpha + \beta x_i)$$

Best to minimise 'sum of squared errors' (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

Take partial derivatives w.r.t. *α* and β to find minimum for each at corresponding values *a* and *b*



$$b = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \qquad a = \overline{y} - b\overline{x} \qquad \Rightarrow \qquad y_{i,\mathrm{mod}} = a + bx_i$$

Errors on linear regression model parameters

If the errors are i.i.d. random variables with the same standard deviation ϵ the errors on the intercept and gradient are normally distributed with the following standard deviations:

$$\mathrm{Err}(a) = \epsilon^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right)$$

$$\mathrm{Err}(b) = \frac{\epsilon^2}{s_x^2}$$

Residuals

data = model + residual

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Now square and sum both sides - terms can also be cancelled since $\sum e_i \to 0$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SST = sum of squared total deviations (scales with total variance)

SSM = sum of squared model deviations

SSE = sum of squared error values

The variance due to the errors, S^2 can be estimated using: $$S^2 = \frac{\text{SSE}}{n-2}$$

Linear regression: caveats

- Takes no account of uncertainty on the *x*-axis values.
- Assumes that the data points are equally-weighted, i.e. the ‘error bars’ on every data point are assumed to be the same.
- Assumes that experimental errors are *uncorrelated* (as with the correlation coefficient)
- Model fitted is linear – this is often not the case but many models may be linearised with a suitable mathematical transformation.
- The same idea of minimising SSE can be applied to non-linear models, but must often be done numerically via computation.
- We will examine much more sophisticated techniques to fit data in a couple of weeks....

Easy error estimation: bootstrapping

- In many cases our data points may not have identical error bars, or for various other reasons it may prove difficult to estimate uncertainties on the model parameters
- A simple but remarkably effective solution is to use the data to do your error estimation for you, using the ***bootstrapping*** approach.

How it works:

If we have *n* data points:

1. Randomly select *with replacement* (i.e. do not take out selected points from the sample – being able to select a data point more than once is crucial for the method!) *n* data points from the sample. For the bivariate case selecting *n* matched pairs of x, y data.
2. Calculate the regression model or other parameters (mean, median, etc.) from the selected data
3. Repeat N_{boot} times, where ideally $N_{\mathrm{boot}} \sim n(\ln(n))^2$
4. Make a distribution of the parameters which you can use to determine confidence intervals etc. (see later)