STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Lecturer: Phil Uttley (p.uttley@uva.nl. C4.145)

TAs: Tom Riley (t.e.riley@uva.nl)
Marieke van Doesburgh
(m.j.vandoesburgh@uva.nl)

How the course works

- **Lecture** (Monday morning): gives key statistical concepts, background theory behind methods & how they work. Slides will be on blackboard, but we also recommend reading the Vaughan course book for deeper info on most topics.
- **Python statistical computing tutorial** (Monday, after lecture): demonstrates the practical side of how to apply statistics to data using Python. You will need to know this in order to do the coursework! The tutorial is done via a jupyter notebook which is available via blackboard, you should bring your own laptops if you have them.
- **Statistical computing workshop** (Wednesday morning): the TAs will be present to help you with your stats programming problems and give advice about the problem sets.

Stats programming and assessment

- We use Python to apply the statistical methods we learn to data (we recommend downloading the Anaconda python package, to be consistent with the modules we use in the tutorials). The computers in the computing Lab also include the Anaconda installation (via Linux).
- There will be plenty of examples given in the tutorials, so even novice Python users should progress quickly, but you may need to put extra effort in (and don't be afraid to ask for help from your colleagues!).
- The assessment of the course is done entirely through problem sets which mostly consist of data analysis problems to be solved using the Python methods discussed in the course (as well as some theory/ mathematical-based questions).
- The problem sets should be submitted via blackboard as jupyter notebooks.

Schedules and assessment

- Teaching schedule:
  - Main lecture: Monday, 9-11h, D1.113
  - Statistical computing tutorial: Monday, 11-13h, F2.04
  - Statistical computing workshop: Wednesday, 11-13h, F2.04

- Assessment:

  5 problem sets to be submitted via blackboard as iPython notebooks, before midnight on the hand-in day (late submission: up to 1 day. -20% of actual marks obtained; up to 2 days: -40%).
  - PS1 (out: today, in: 7/11) 15% of total grade
  - PS2 (out: 7/11, in: 14/11) 15% of total grade
  - PS3 (out: 14/11, in: 21/11) 15% of total grade
  - PS4 (out: 21/11, in: 5/12) 20% of total grade
  - PS5 (out: 5/12, in: 19/12) 35% of total grade

Course outline

Core material mostly follows the course book, “Scientific Inference: Learning from Data” by Simon Vaughan (Cambridge University Press), e.g. available from bol.com or amazon.de

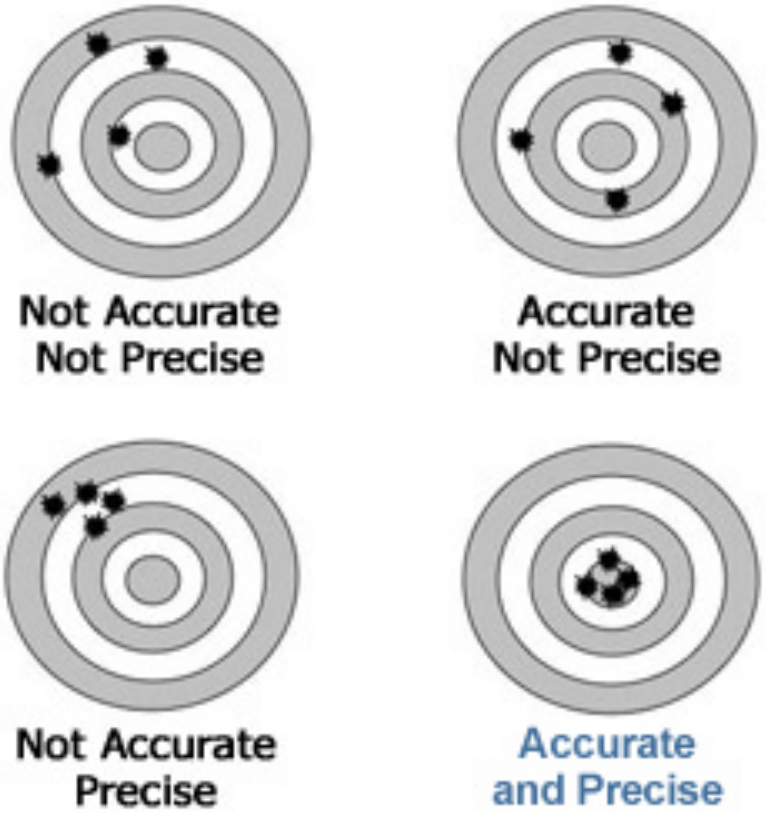
- Week 1: Statistical summaries of data and simple statistical inference (chapters 1-3)
- Week 2: More on correlations. Probability theory (chapter 4)
- Week 3: Random variables (chapter 5)
- Week 4: Model fitting: Estimation and maximum likelihood (chapter 6)
- Week 5: Model fitting: Significance tests and confidence intervals (chapter 7)
- Week 6: Hypothesis testing and likelihood ratio (additional material)
- Week 7: Monte Carlo methods (chapter 8)

STATISTICAL METHODS FOR THE PHYSICAL SCIENCES

Week 1: Statistical summaries of data and simple statistical inference

Basics 1: errors, precision and accuracy

- We need statistics because our data is affected by random error.
- Measurements are randomly *sampled* from a *population*. This could be a real underlying population or a notional population of 'possible' measurements.
- The *statistical error* depends on the properties of this underlying population (the statistical *distribution*) as well as the sample size.
- Precision is related to the size of the statistical error.
- *Accuracy* is limited by *systematic errors* or *biases*. Maybe there is something wrong with our measuring apparatus or its calibration. Or there is a bias in our sampling approach.



Basics 2: data

What type?

- categorical, ordinal, discrete, continuous
  - in the physical sciences we mostly consider the last two types, and some categorical, we mostly leave ordinal data to the social scientists!

How many dimensions?

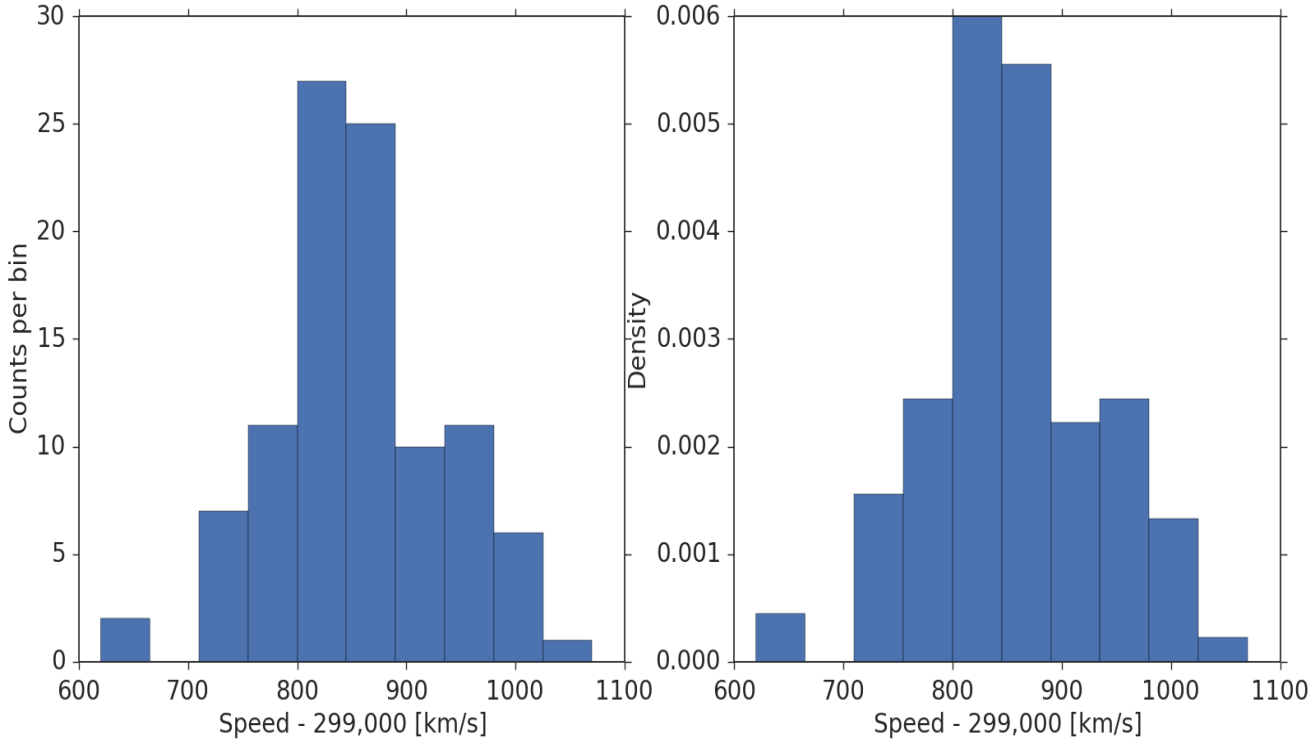
- univariate, bivariate, multivariate

Variables

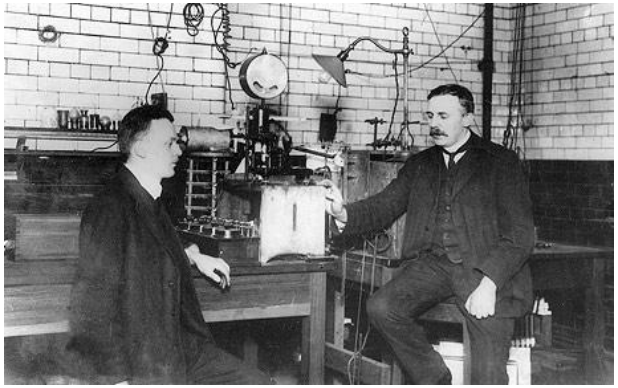
- explanatory/independent, response/dependent
  - why might ‘explanatory’ and ‘response’ be better terms than ‘independent’ and ‘dependent’?

Plotting continuous univariate data: histogram of Michelson's measurements of c

Albert A. Michelson

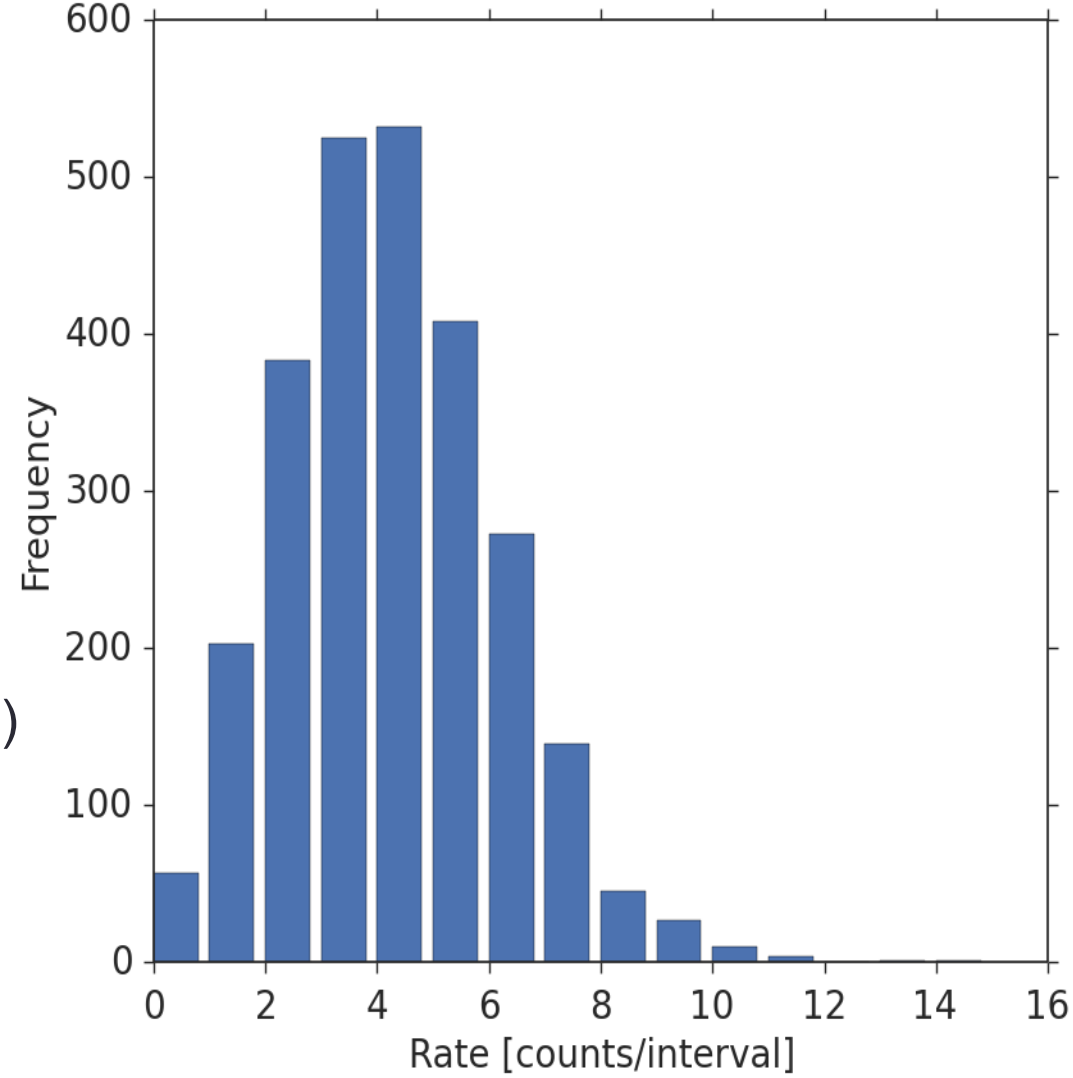


Plotting discrete/categorical univariate data: the bar chart

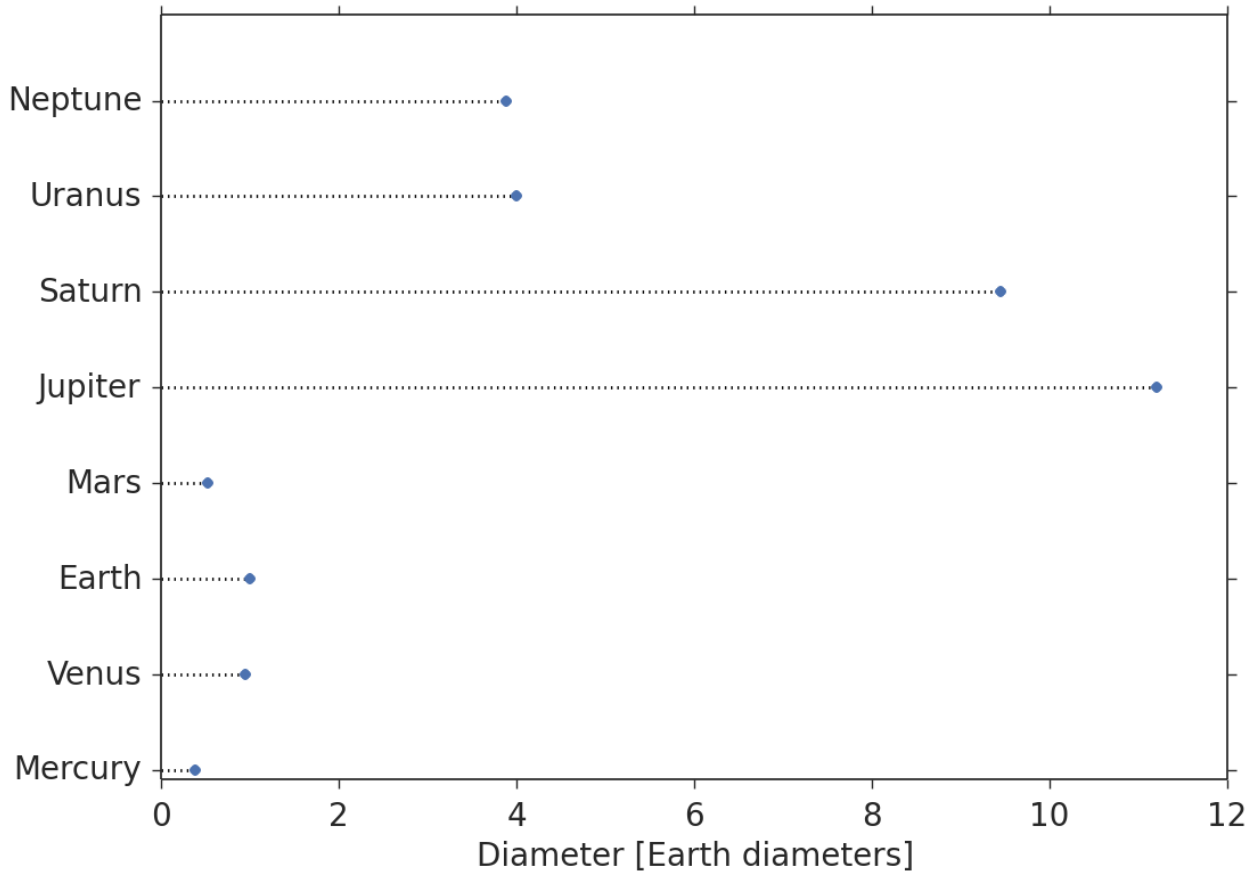


Rutherford and Geiger's data (scintillations associated with α-particles from polonium decay, counted in intervals of 7.5s)

What type of distribution is this?



Categorical data: dot charts



- Dot charts are a useful way to compare categorical data. They can also be used to show the distributions of individual measurements in different samples (categories)

The ‘centre’ of data: mean, median and mode of a sample (x) of n values.

- The mean of x:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The median of x:
  - Sort the values of x into ascending order. The median is the ‘middle’ value [(n+1)/2]th value if n is odd or the average of [n/2]th and [$n/2$+1]th values if n is even.

- The mode of x:
  - The value of x with the most measured values (maximum in a histogram or probability distribution)

Dispersion in data: variance and standard deviation

- The variance is always a positive quantity:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

why not normalise by n? 'Bessel's correction' – see Box 2.2 of Vaughan

- Another form, expanding terms (angle brackets also denotes averaging):

$$s_x^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2\right) = \frac{n}{n-1}\left(\left\langle x^2\right\rangle - \left\langle x\right\rangle^2\right)$$

called a 'mean-squared' value, meaning that it is the mean of something squared, not the square of the mean!

- Standard deviation is the square-root of variance. It is in some sense the 'spread' in the data. It is sometimes called the rms ('root mean squared') deviation, or 'sigma' (σ). The latter term means something specific in statistics (related to normal distributions) so we will avoid it here.

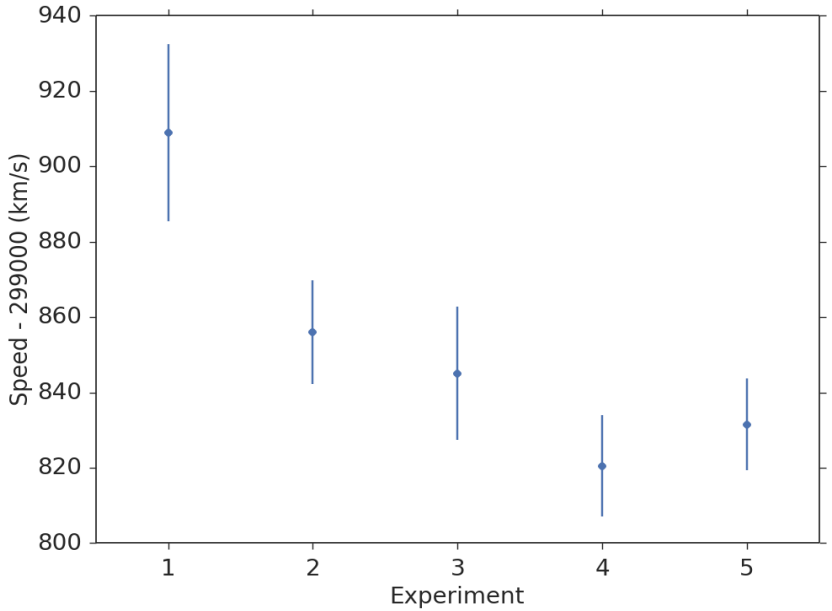
Standard errors and error bars

The standard deviation tells us the spread of values, but how accurately is the mean determined?

The *standard error* on the sample mean is: $SE_{\bar{x}} = \sqrt{\frac{s_x^2}{n}}$

The standard error can be used as an 'error bar' on the mean.

Comparison of mean values of *c* (with errors) for each of Michelson's experiments



But what does this actually tell us? For now we will assume it gives the 'expected' uncertainty on the true value of the mean of the underlying population of values of *x*.

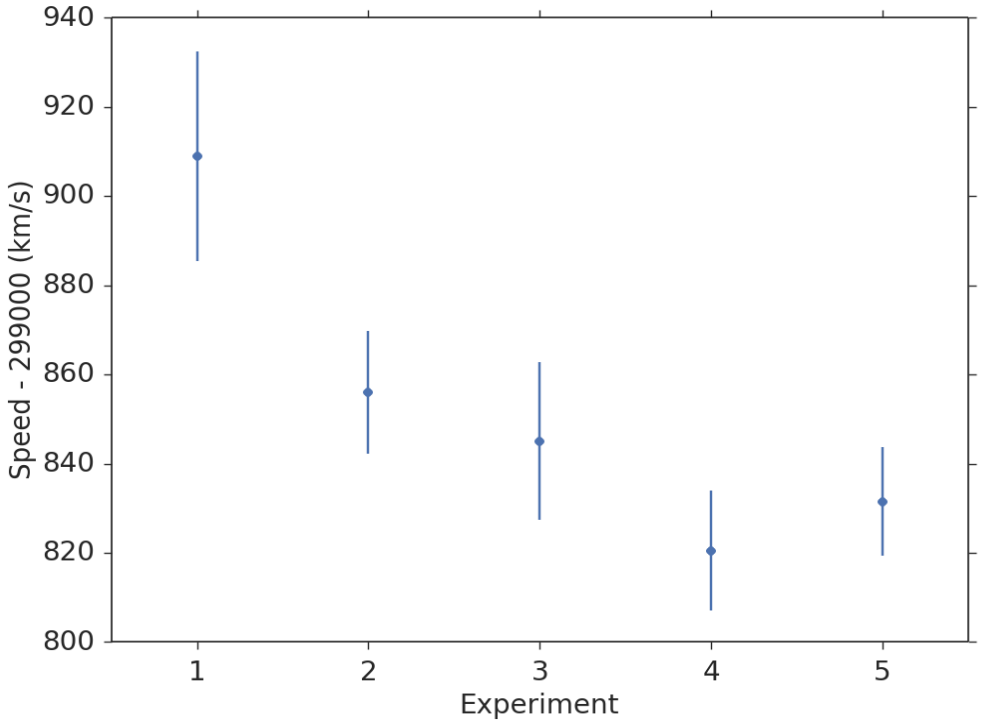
Deviations from the mean

The *standard error* on the sample mean is:

$$SE_{\bar{x}} = \sqrt{\frac{s_x^2}{n}}$$

The standard error can be used as an ‘error bar’ on the mean.

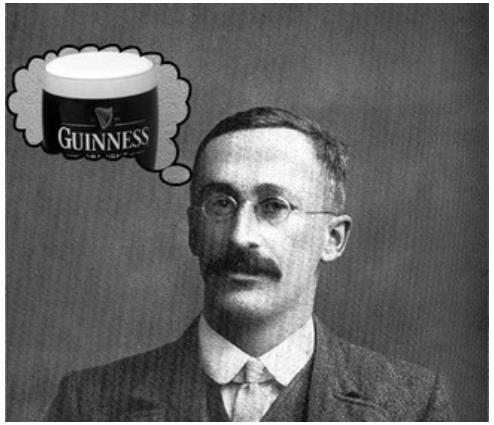
Comparison of mean values of *c* (with errors) for each of Michelson’s experiments



How big a deviation? Student's *t*-statistic

- If we already know the value (μ) we are comparing with, we can use the one-sample *t*-statistic:

$$t = \frac{\text{observed difference}}{\text{standard error}} = \frac{\bar{x} - \mu}{\sqrt{s_x^2 / n}}$$



William Sealy Gossett, aka "Student"

- Now we measure means from two separate samples and compare, we use the two-sample *t*-statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_{x1}^2 / n + s_{x2}^2 / n}}$$

Errors add in quadrature

- The statistic gives an indication of how likely the means are to be drawn from the same population (i.e. same underlying mean).
- If the errors are normally distributed then in certain circumstances that must be specified, e.g. when the samples have the same number of measurements and are drawn from populations with the same variance, the t-statistic can be used to give a precise probability that the population means are the same (see later).

Quantiles of a sample

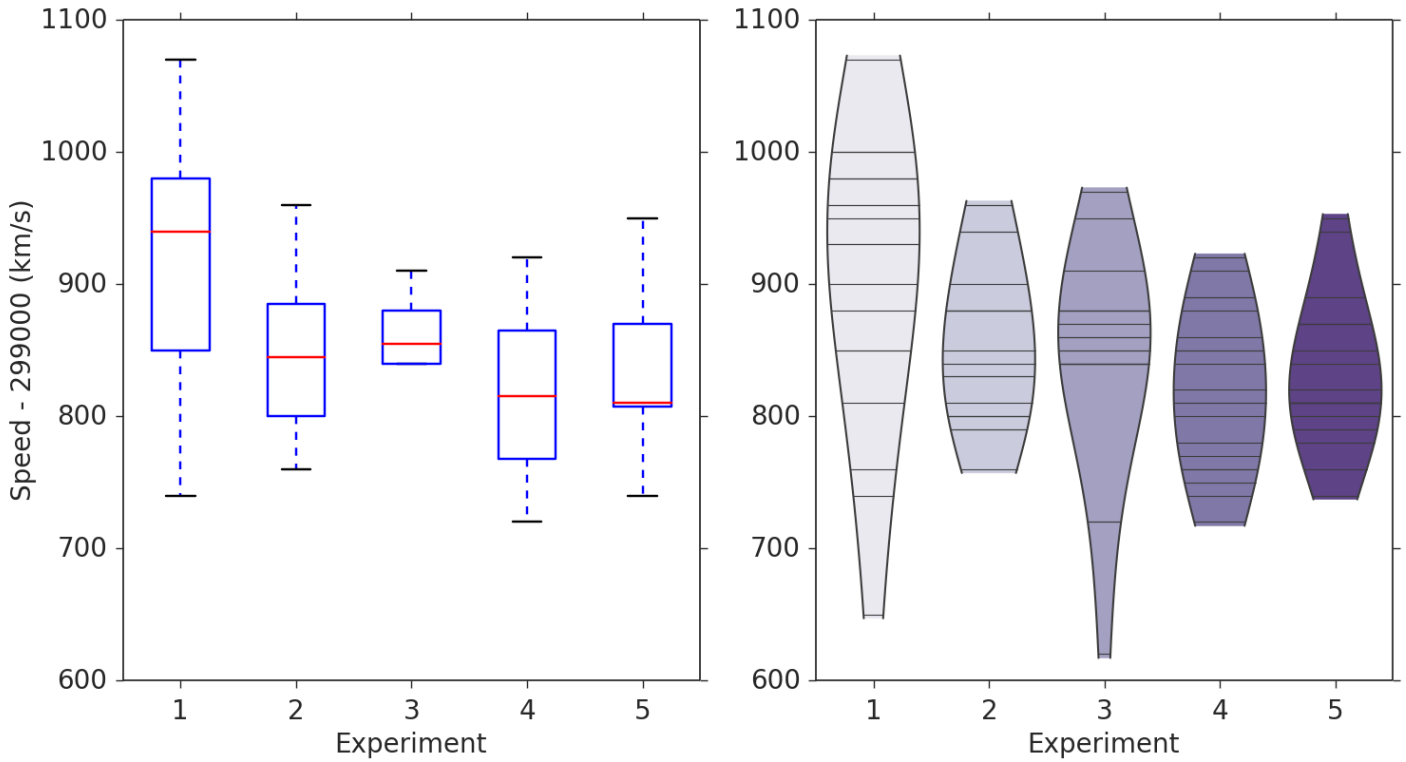
- In a sample ordered according to *x, t*he *α* quantile is the value of *x* below which a fraction *α* of the sample can be found.

- Special names are given to these quantiles: 0.25 (first quartile), 0.5 (second quartile or median) and 0.75 (third quartile). The difference between the first and third quartiles is called the *interquartile range (IQR).*

- Can also be expressed as a percentage (called percentile, e.g. the first quartile is the 25^{th} percentile.

Box plots and violin plots

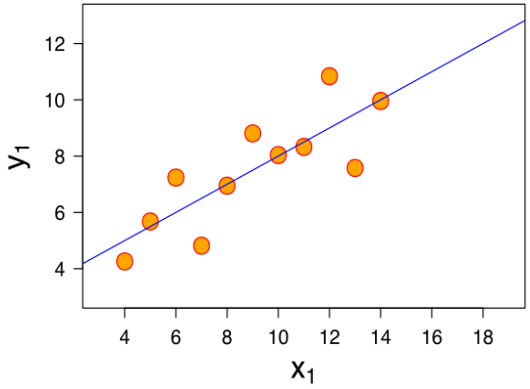
- The box plot shows 1.5×IQR (the ‘whiskers’), and the 1st, 2nd and 3rd quartiles.
- The violin plot shows individual values (horizontal lines) and uses a ‘kernel density estimate’ to estimate the density of values as a smooth function, which changes ‘width’ of the violin, giving an indication of the probability density function of the points.

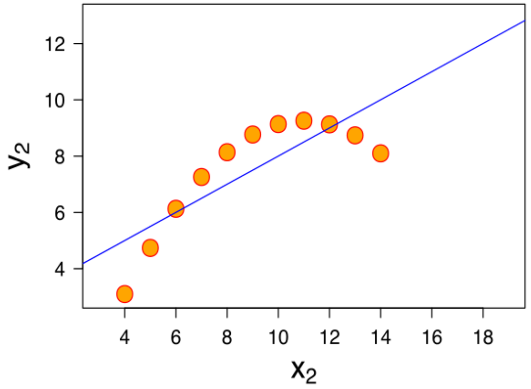


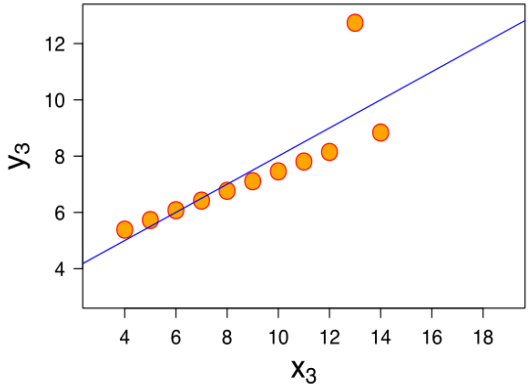
Bivariate data: scatter plots

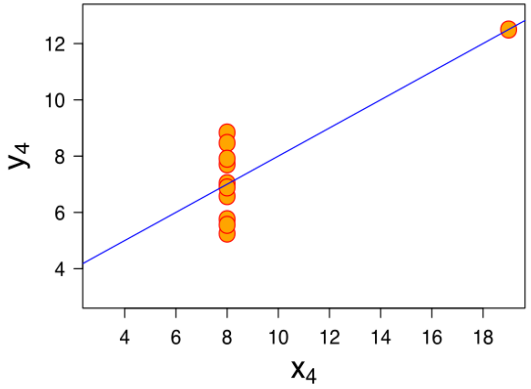
- We can plot the '*response'* y variable versus variable x ('e***x***planatory')
- Plotting reveals much more than simple statistical measures!
- Always plot your data!

'Anscombe's quartet': all have the same x and y mean and variance and the same correlation coefficient and best-fitting linear regression model (see later)

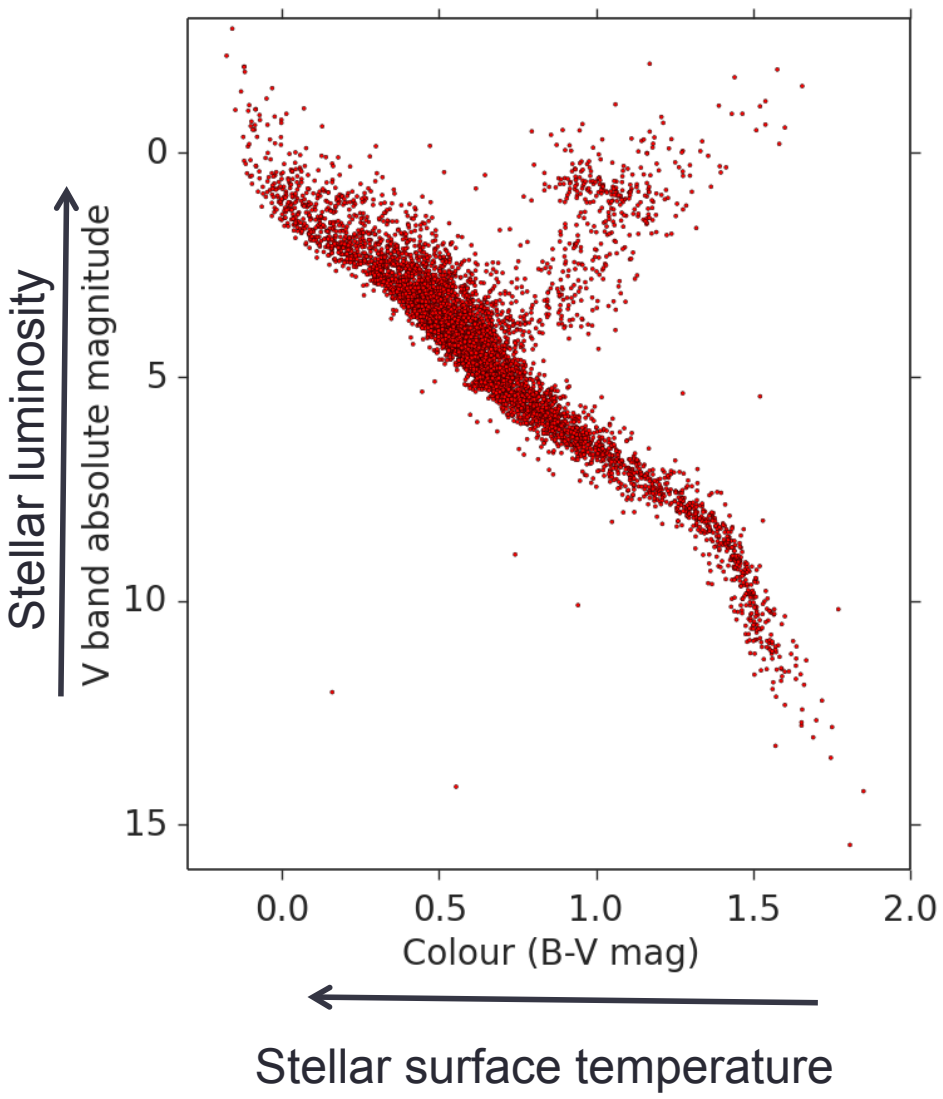


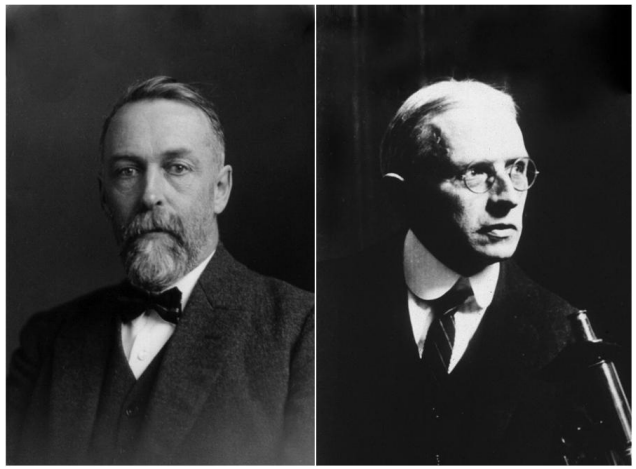






The Hertzsprung-Russell Diagram





Ejnar Hertzsprung

Henry Norris Russell

Hipparcos

The sample covariance and correlation coefficient

- We can determine statistical properties for each variable (e.g. mean, variance) separately, but this does not tell us if (and how) they are related.
- To do this we can define the sample covariance, s_{xy}:

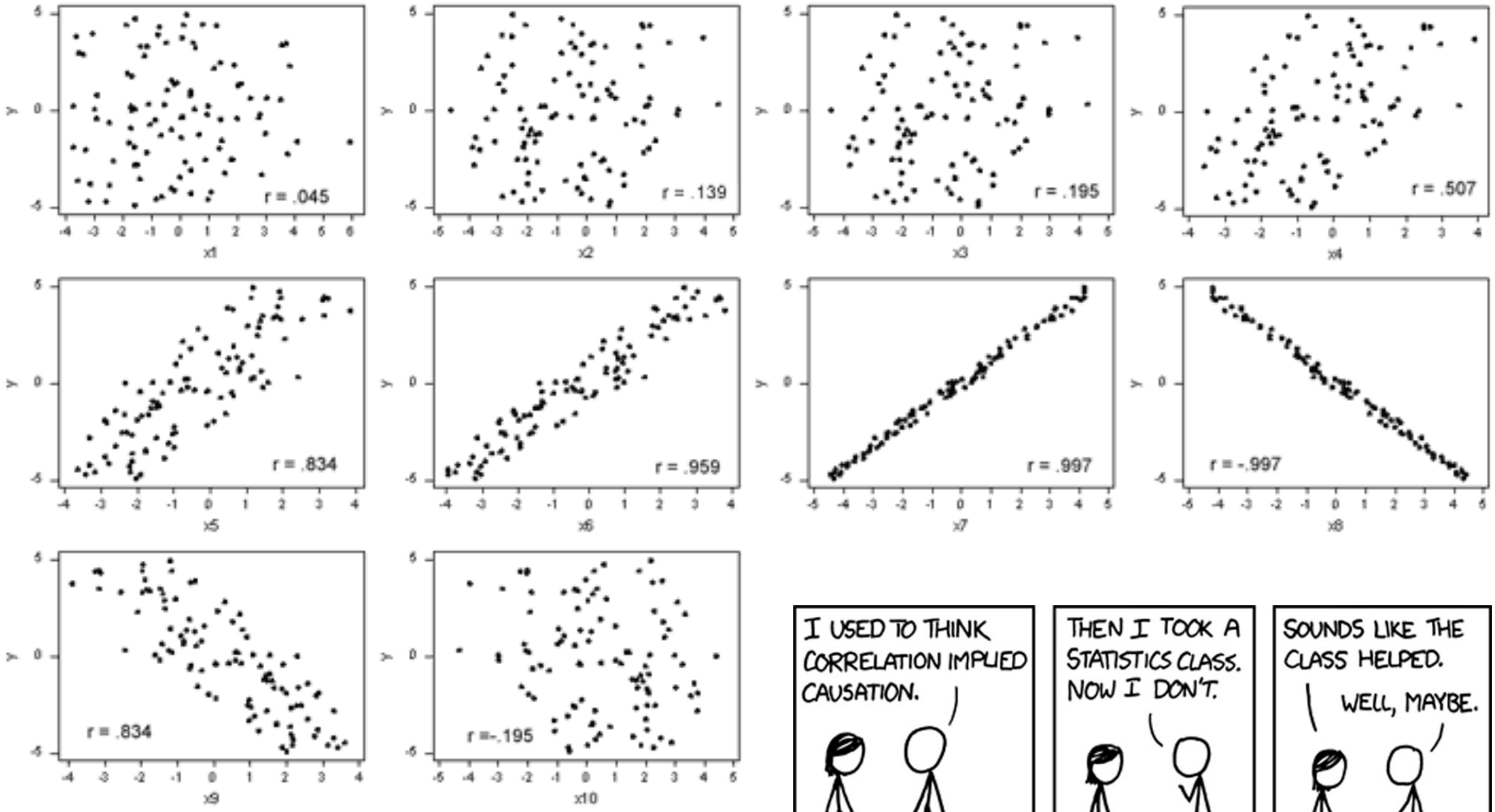
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

- Note that s_{xx} is just the sample variance!
- We can normalise by the standard deviations to obtain the *correlation coefficient, r* :

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

- *r* gives us a way to compare the correlations for variables with very different magnitudes. In some (very specific) circumstances it also translates into a probability that the correlation is statistically significant (more on this later).

Example correlations

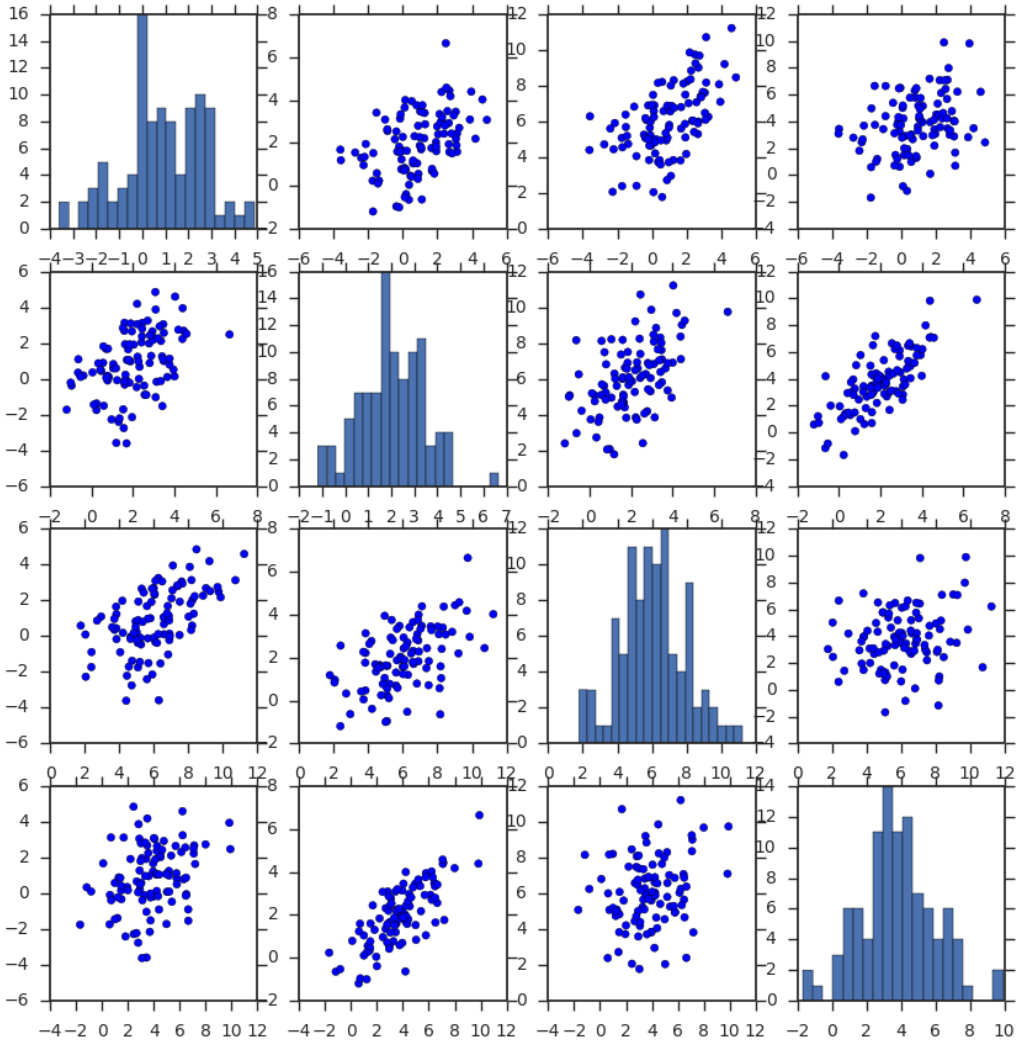


(xkcd.com)

Remember, correlation is not causation!!! (more on this later)

Plotting multivariate data

- It can be difficult to convey multivariate information clearly.
- A third parameter could be represented by a contour or density plot, or even 3D graphics.
- A matrix of scatter plots can be useful to examine correlations between many parameters



The diagonals show histograms (otherwise they would plot the same variable against itself!)

Good practice in (statistical) graphics

- Show your data, without distortion, with clear labelling (and/or clear caption). A good principle: can you port the figure straight to a talk slide, such that the audience can clearly see the labels and data-points/lines (are they too small/faint?).
- Use a plot appropriate for the data – what are you trying to show?
- Try to show the greatest amount of information ***as clearly as possible*** in the available space.
- Use colour carefully (remember some-people are colour-blind!). You can use different marker styles to enhance the differences, but make sure they are visible and do not clutter the plot too much!
- Use multipanel plots if necessary to convey information clearly.
- Put yourself in the shoes of a new reader unfamiliar with your work. Ask a colleague if the plot makes sense to them!