Basic Statistics in Bibliometrics

Adam Finch | Analyst, CSIRO
Statistics

- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

- Uses a number of complicated formulae and calculations to interpret data, based on huge tables of cross-referenced values

- Can look at whether something is a real effect or just an apparent effect, whether one thing explains another, how different something is from something else

- Full of misleading terms like ‘significance’ and competing schools of thought

- A dark art more than a science as it requires human decisions and varies with field

- You have to know which analysis to use in which situation, otherwise the results are wholly misleading

- In research, relies strongly on experimental design; in bibliometrics, stats are a minefield...
Causation vs Correlation

- Examples: Dawn Chorus and International Talk Like a Pirate Day

- Statistics often attempt to strip out any other influences ("confounding variables") to demonstrate that x is a cause of y

- Be wary of jumping to conclusions...

- Does Open Access cause higher citations?

- Do US authors like to collaborate more than Singaporean authors?

- Does publication in *Nature* cause higher citation?

- Are English-speaking nations better at science?
Sample Sizes

• Significance

• Crucial to bear in mind for individual researchers, teams, small subjects...

• When data is normally distributed, you need around 30 data points for a reliable mean (p < 0.05)
Options for Statistical Analysis

• Don’t do it!

• Test whether data normally distributed then hope for the best!

• Transform data – take square root or log transformation

• Use non-parametric alternatives e.g. Mann-Whitney test, Krustal-Wallace test or Wilcoxon signed-rank test

• [http://www.real-statistics.com](http://www.real-statistics.com)