Quantifying Uncertainty:

Estimates and Confidence Intervals

Statistics for Responsible Metrics
Session Three

Copenhagen University Library, Denmark, 12th December 2018
Dr Abigail McBirnie
In bibliometrics....

1. Are publication counts estimates?
2. Are citation counts estimates?
3. Are normalised indicators estimates?

Yes, all bibliometric indicators are estimates
But, estimates of what?

- publication counts = estimate of publications → not all publications may be in the database and there may be indexing errors

- citation counts = estimate of citations → not all citing publications may be in the database and there may be missing citation links and other errors

- normalised indicators are estimates → they are calculated from indicators that are themselves estimates (e.g. publication and citation counts)
Other points to consider:

Random and social factors affect both publication and citation.

During periods of change (e.g. moving from one field to another) there may be instability that affects both publication and citation.

Bibliometric indicators are not estimates of quality, impact, or capability.
Quantifying uncertainty

As an estimate, a bibliometric indicator could reasonably take other values within a certain range. This could happen randomly or due to small changes in the dataset or errors.

Formulae and bootstrapping can be used to estimate the range that an indicator could reasonably have taken.

These approaches all assume the ‘no change hypothesis’ → that things continue as they are.
A confidence interval

- assumes the no change hypothesis
- is reported at a certain level (e.g. 90%, 95%, 99%)
- gives a range of expected values (an interval)
- thus, a 95% confidence interval gives the range of expected values that should occur 95% of the time under the no change hypothesis
In bibliometrics, confidence intervals are often more formally called ‘stability intervals’

The key point here is that any one of the expected values within the confidence interval could be a plausible result for the indicator—not just the particular actual result (single point estimate) you calculated.
I’ve heard of ‘standard deviation’...

- **Variance** is a measure of the variability of your data.

- Reported in units\(^2\): \[ s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \]

- Difficult to interpret units\(^2\) so we take the square root = **standard deviation** \( s \)

- Rule of thumb: for normally distributed data, 95% of the values should be within 2 standard deviations of the mean.

- In bibliometrics, standard deviation may be useful in formulae for confidence intervals, but not on its own.
 Parametric confidence intervals

- For data that are approximately normally distributed, the 95% confidence interval for the population mean is:

\[ \bar{x} \pm t_{0.05,n-1} \frac{s}{\sqrt{n}} \]

- \( t_{0.05,n-1} \) is a Student t value that can be looked up in a table.
- Population mean \( \rightarrow \) the long term mean under the no change hypothesis.
- Calculate sample mean and sample deviation and then apply the above formula.
R Example of Parametric CI

```r
> y <- rnorm(1000, 1)
> stat.desc(y, norm=T)

<table>
<thead>
<tr>
<th>nbr.val</th>
<th>nbr.null</th>
<th>nbr.na</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000.00000000 0.00000000</td>
<td>0.00000000</td>
<td>-2.14724561</td>
<td>3.80193526</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>range</th>
<th>sum</th>
<th>median</th>
<th>mean</th>
<th>SE.mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.94918087 1006.18635226</td>
<td>1.00361864</td>
<td>1.00618635</td>
<td>0.03070488</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CI.mean.0.95</th>
<th>var</th>
<th>std.dev</th>
<th>coef.var</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.06025347   0.94278987</td>
<td>0.97097367</td>
<td>0.96500382</td>
<td>0.05936272</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>skew.2SE</th>
<th>kurtosis</th>
<th>kurt.2SE</th>
<th>normtest.W</th>
<th>normtest.p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.38375870</td>
<td>-0.11243723</td>
<td>-0.36379387</td>
<td>0.99881758</td>
<td>0.76624487</td>
</tr>
</tbody>
</table>

> t.test(y)

One Sample t-test

data:  y
t = 32.77, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  0.9459329 1.0664398
sample estimates:
mean of x
1.006186
But bibliometric data are often non-normal...

One solution is log transformation to normalise the data

→ add a constant to each data point, and then take the log of each data point

✔ solution suggested by Thelwall → MNLCS

see: https://arxiv.org/abs/1612.01431
R Example of Log Transformation

```r
x<-c(0,0,9,4,2,52)
logx<-log(1+x)
logx
[1] 0.000000 0.000000 2.302585 1.609438 1.098612 3.970292
stat.desc(logx, norm=T)

  nbr.val      nbr.null      nbr.na       min      max       range
  6.0000000   2.0000000   0.0000000   0.0000000 3.9702919 3.9702919

  sum     median      mean SE.mean CI.mean.0.95     var
  8.9809272 1.3540251 1.4968212 0.6169687 1.5859687 2.2839026

  std.dev coef.var skewness skew.2SE kurtosis kurt.2SE
  1.5112586 1.0096454 0.4291330 0.2538785 -1.4689747 -0.4219308

  normtest.W normtest.p
  0.9216702 0.5174709
```

```r
stat.desc(x, norm=T)

  nbr.val      nbr.null      nbr.na       min      max       range
  6.0000000   2.0000000   0.0000000   0.0000000 52.0000000 52.0000000

  sum     median      mean SE.mean CI.mean.0.95     var
  67.0000000 3.0000000 11.1666667 8.28016371 21.28483844 411.3666667

  std.dev coef.var skewness skew.2SE kurtosis kurt.2SE
  20.28217608 1.81631428 1.28146139 0.75812278 -0.22169473 -0.06367696

  normtest.W normtest.p
  0.63491446 0.00118412
```
Limitations of parametric confidence intervals

- In bibliometrics, we often have non-normal data.
- Parametric confidence intervals will not work where there is skew → log transformation can help.
- Parametric confidence intervals require that data are independent from each other.
- But, with publications, related publications may tend to be cited together, i.e. we cannot say the citations are independent of each other.
- So confidence interval = stability interval, which is only an approximation.
Non-parametric confidence intervals

To calculate confidence intervals, we can use approaches that are based on known but non-normal distributions

- **Wilson score interval** for success/failure type data → this uses the binomial distribution

- Confidence intervals for count data using the Poisson distribution

We can also use **bootstrapping**, which is a ‘distribution-free’ approach → we do not need to make any assumptions about the distribution of our data
What is bootstrapping?

• Useful if data do not follow a standard statistical distribution

• How it works:
  ➢ The sample is treated as the population
  ➢ Resample from this 1000+ times
  ➢ Calculate the indicator of interest for each of these 1000+ resampled samples
  ➢ Let us say you want a 95% confidence interval
  ➢ Determine which of your 1000+ results is the 2.5%th smallest indicator
  ➢ Determine which of your 1000+ results is the 97.5.5%th largest indicator
  ➢ Use these two results as your confidence interval
Bootstrapping Example

Original sample: 0, 1, 2, 3, 5, 7, 11, 13, 25 → mean=7.4

- Resample 1: 0, 0, 2, 3, 3, 7, 11, 13, 13 → mean=5.8
- Resample 2: 0, 0, 0, 0, 2, 7, 11, 13, 25 → mean=6.4
- Resample 3: 2, 2, 2, 3, 3, 7, 11, 13, 25 → mean=7.6
- ...
- Resample 1000: 0, 0, 2, 3, 3, 7, 11, 13, 13 → mean=5.8

The 25th smallest is 4.9 and the 975th largest is 7.5 so the bootstrapped 95% confidence interval is (4.9, 7.5) around the original sample calculated mean of 7.4

Note the asymmetry possible in bootstrapped confidence intervals!
Limitations of non parametric confidence intervals

- Even if a known non-normal distribution is used, your data may still only approximate the selected distribution.
- Bootstrapping becomes less reliable the smaller the sample, especially with highly-skewed data.
- As with parametric confidence intervals, you are still making assumptions that the data are independent of each other, which may not necessarily be the case with citations.
- So confidence interval = stability interval → treat as an approximation.
For a useful summary of the topic of confidence intervals, see: