I. Visualising Uncertainty in Bibliometric Data

II. Key Issues to Consider when Inferring Difference

Statistics for Responsible Metrics
Session Four

Copenhagen University Library, Denmark, 12th December 2018
Dr Abigail McBirnie
Part One:

Visualising Uncertainty in Bibliometric Data
Ludo Waltman: Six principles for scientometric visualisation

- Acknowledge different use cases of visualisation
- Adjust visualisations to the mode of presentation
- Find an appropriate balance between technical sophistication and methodological transparency
- Be aware that visualisations tend to give a simplified and incomplete representation of the underlying data
- Combine visualisations with other pieces of evidence
- Test sensitivity of visualisations to methodological choices and handle these choices pragmatically

Principle 1: Acknowledge different use cases (i.e. providing insight versus attracting attention)

If your focus could only be on one of these, which would it be?

Focus of visualisation for responsible metrics → providing insight

What is the focus of visualisation for many vendor tools?
- quick wins (both for you as an analyst and as a customer)
- ease of use (range of users, not all experts)
- computing speed and ability to handle very large datasets
- accuracy and coverage (??)

As such, I would argue that the focus is on → attracting attention

This is potentially problematic for our goal of responsible metrics
Principles 4 and 5:

- Visualisations are simplified and incomplete representations
- Combine visualisations with other pieces of evidence

Key questions to consider:

1. In what ways are our visualisations simplified/incomplete?
2. What interpretation gaps do we need to fill?
3. How can we do this?
4. What evidence can we add?
5. What are the implications for responsible metrics?
Example One: Visualisations that better illustrate the nature of bibliometric distributions \(\rightarrow\) skew

The normal distribution is our default distribution ‘picture’ \(\rightarrow\) we need to emphasise the skewed nature of bibliometric data

Lariviere et al (2016) propose accompanying JIFs with distributions of the underlying data in order to illustrate the skew

In turn, Phil Davis from thescholarlykitchen blog replies, including this image:

The blog post’s main point is that the JIF is a problematic indicator, even with the accompanying citation distribution visualisations. I think we would agree!

Nonetheless, could such visualisations help to fill an interpretation gap? What additional evidence does the above image incorporate?

Example Two: Visualisations that better illustrate the nature of bibliometric distributions → discrete data

Our default distribution ‘picture’ imagines continuous data. We need to emphasise that most bibliometric data is discrete.

Possible action: Avoid implying undue precision. Consider rounding

Possible action: Use discrete histograms, where appropriate, instead of continuous ones

FWCI = 1.28

SciVal could show this as:

FWCI = 1.3 or FWCI = 1.28
Histograms in R

\[ x <- c(0,0,0,0,1,1,2,3,3,4,4,10,22) \]

**Continuous histogram:**  
\[ \text{hist}(x, 7) \]

**Discrete histogram:**  
\[ \text{plot(table}(x), \text{type}="h", \text{lwd}=5, \text{ylab}="Freq")} \]
Example Three: Visualisations that illustrate uncertainty for an indicator → include confidence intervals

The Leiden Ranking has done this for some years:

Example visualising confidence intervals in R
Principle 6: Visualisations and methodological choices

Visualisations can provide shortcut methods, such as in the case of ‘eyeballing’ confidence intervals

When eyeballing confidence intervals, how data are visualised matters...

Compare...

With ...

But be careful with such approaches....

Your eye may not be the best choice of measuring tool.

In bibliometrics, eyeballing confidence intervals is the most conservative approach. Hypothesis testing may suggest a difference that your eye cannot see.
Part Two:
Inferring Difference in Bibliometrics
Consider three key statistical issues:

1. Hypothesis testing
2. Statistical power
3. Adjusting for multiple comparisons
1. Hypothesis testing
Hypothesis Testing

• Hypothesis → an assumption about something

• Hypothesis test → a statistical procedure that indicates whether or not we can accept our assumption

• Our null hypothesis ($H_0$) is that the two indicators are randomly drawn from populations with the same underlying value (e.g. mean)

• Our alternative hypothesis ($H_1$) is that the two indicators are randomly drawn from populations with a different underlying value

• We can test our hypotheses at different levels of confidence (e.g. 0.05, 0.01, 0.01) → p-values

• The test suggests whether or not there is enough evidence to conclude that the difference in observed values is likely to reflect an underlying difference

• In practice, the result is a yes/no type decision
Why use hypothesis testing?

First, a disclaimer....

In statistical terms, it is slightly more complicated than what we’ve discussed the precise interpretation of $H_0$ and $H_1$ depends on which statistical paradigm you follow and which assumptions you make.

For example, you could test not just whether or not a value is different, but rather, whether or not it is larger (or smaller).

Essentially though, the purpose of all hypothesis testing is the same:

***To reduce the chance of incorrectly concluding that one group is better than another based on a minor difference between indicators***

Because hypothesis testing is a formal statistical procedure, it is more sensitive than eyeballing confidence intervals.
Hypothesis Testing Example

✓ University A has a CNCl of 1.14
✓ University B has a CNCl of 0.94
✓ A hypothesis test of difference reports p=0.02.
✓ Since p < 0.05 we reject the null hypothesis at the 0.05 level
✓ Since p ≥ 0.01 we retain the null hypothesis at the 0.01 level
✓ We can report this as “some evidence” to reject the null hypothesis and retain the alternative hypothesis.
✓ It is common to report the strongest level at which the null hypothesis is rejected (0.05, 0.01 or 0.001)
✓ For decision making, should decide before testing which level to use
Example R Commands for Hypothesis Tests

Two important points:
1. These t-tests assume your data are normally distributed.
2. You would usually only compare (i.e. whether or not there is a difference) if you are dealing with normalised bibliometric indicators.

`t.test(x,y, var.equal=FALSE, paired=FALSE)`
→ Use to test two groups
→ x and y are vectors of the data for the two groups

`t.test(x,mu=1.00)`
→ Use to test one group against a known value (e.g. the world average of 1.00)
→ x is a vector of the data for your group
Comparing the means of two samples

```r
> x<-rnorm(5000)
> y<-rnorm(6000)
> t.test(x,y,var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test

data:  x and y
t = 0.47484, df = 10653, p-value = 0.6349
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.02835939  0.04649162
sample estimates:
  mean of x  mean of y
0.018049460  0.008983344
```

Comparing your sample mean with a known value (e.g. world average 1.00)

```r
> z<-rnorm(1000,5)
> t.test(z, mu=1.00)

One Sample t-test

data:  z
t = 129.62, df = 999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
  4.966440  5.088387
sample estimates:
  mean of x
5.027413
```
Type I and II Errors

- A **Type I** error is when you **incorrectly reject the null** hypothesis
  
  (In other words, you say there is a difference when there isn’t)

- A **Type II** error is when you **incorrectly retain the null** hypothesis

  (Here, you fail to spot a difference when there is one)

- Generally, the concern is more with Type I errors

- We can see this with the focus on p-values → p<0.05 means you would expect to make a Type I error 5 out of 100 times; p<0.01 means you would expect to make that error 1 in 100 times; etc.
Limitations of Hypothesis Testing

Hypothesis testing → **whether or not** there is a difference

Confidence intervals → **how much** of a difference

Hypothesis testing:

- Binary and has a cut off point (i.e. you lose information)
- t-test relies on having normally distributed data
- In a bibliometric context, assumes certain conditions exist: no change and same capability
- Requires care about p-value decisions when multiple comparisons are made
- Differences may be statistically significant but not relevant for practice → this can happen where very large datasets are involved, e.g. UK cancer research is 0.01% more cited than US cancer research
2. Statistical power
What is statistical power?

- The **power** of a hypothesis test is the probability that the test rejects the null hypothesis ($H_0$) when the alternative hypothesis ($H_1$) is true.

- Statistical power calculations tell us **how much data to collect** in order to have confidence that a moderate difference can be detected if it exists.

- Power should be **calculated before the study begins**.

- There are **mathematical formulae** for calculating power→ these are based on assumptions of data distributions and useful minimum difference.
How is statistical power relevant for bibliometrics?

- In bibliometrics, we often have a fixed amount of data → e.g. the publications of a research group found in Web of Science for the period (2012 – 2015)

- Usually, the only way to increase the sample size is to expand the publication window → e.g. widen the window from (2012 – 2015) to (2010 – 2015)

- But, we might have real-world, practical reasons for not wanting to do this

- Nonetheless, it would still be useful for us to know:
  1. if we had roughly enough data
  2. how ‘close’ we were to having enough data
**Post hoc power** is statistical power that is calculated after the study is complete → formal calculation procedures exist.

Unfortunately, the calculations involved create an undesirable circular argument → as such, many argue ‘**Do not use**’.
But, remember, we still want to know:

1. if we had roughly enough data
2. how ‘close’ we were to having enough data

As an informal approach, eyeballing confidence intervals can help

We can look at:

1. the width of the confidence interval
2. how tightly focused the confidence interval is on zero (or another value of interest)
Lots of data...

```r
> x<-rnorm(5000)
> y<-rnorm(6000)
> t.test(x,y, var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test

data:  x and y
t = 0.47484, df = 10653, p-value = 0.6349
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.02835939  0.04649162
sample estimates:
  mean of x  mean of y
  0.018049460  0.008983344
```

Eyeballing Power Example 1

Only a little data...

```r
> x<-rnorm(5)
> y<-rnorm(6)
> t.test(x,y, var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test

data:  x and y
t = 0.15384, df = 5.3275, p-value = 0.8834
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.283900  1.450584
sample estimates:
  mean of x  mean of y
-0.05525848 -0.13860017
```

Narrow confidence interval, closely around 0

Still contains 0, but very wide interval
Lots of data...

```r
> z<-rnorm(1000,1)
> t.test(z, mu=1.00)
```

```
One Sample t-test

data:  z
t = 0.83512, df = 999, p-value = 0.4039
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
  0.9644152 1.0883119
sample estimates:
mean of x
1.026364
```

Eyeballing Power Example 2

Only a little data...

```r
> z<-rnorm(10,1)
> t.test(z, mu=1.00)
```

```
One Sample t-test

data:  z
t = -0.99968, df = 9, p-value = 0.3436
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
  0.05004258 1.36767602
sample estimates:
mean of x
0.7088593
```

Narrow confidence interval, closely around 1

Still contains 1, but a much wider interval
Remember that statistical power relates to **Type II errors**: Considering statistical power is useful when you get an insignificant result in a study and wonder if this could be down to not having enough data

But, be careful about the conclusions you draw...

Would having enough data next time lead to a significant result?

→ Statistical power **cannot** tell you this.
3. Adjusting for multiple comparisons
Do we make multiple comparisons in bibliometrics?

• Yes → we compare different groups

• Yes → good practice is to use a ‘basket of metrics’ for a group

• However, this creates complications when we want to use statistical inference

• Why? → for mathematical reasons, our chance of making a Type I error (incorrectly rejecting the null hypothesis) increases when we make multiple comparisons
A multiple comparison example...

- You are looking at a bibliometric indicator (e.g. FWCI) for 5 universities

- You run a series of hypothesis tests → you want to see if each of these 5 indicators differ from the ‘norm’

- You now have 5 resulting p-values, 1 of which is significant at the 0.05 level

- Remember we said earlier that (assuming certain conditions) we could interpret a p-value of 0.05 as saying that we would wrongly conclude that there was a difference 5 out of 100 times? Because of our multiple comparisons here, our chance of doing this is actually much higher

- Mathematically, if there are no differences, then the chance of rejecting all 5 null hypotheses is: \(0.95 \times 0.95 \times 0.95 \times 0.95 \times 0.95 = 0.95^5 = 0.77\) (77%)

- So there is actually a 23% chance we will make a Type I error, not a 5% chance

- And, if we were comparing 14 bibliometric indicators, then the chance of making a Type I error would be 51% → in other words, we would be more likely than not to make a mistake!
How to deal with this problem?

- You can apply an adjustment calculation for multiple comparisons

- Examples include:
  
  **Benjamini-Hochberg** → use when you are comparing an indicator calculated from multiple, different datasets

  **Benjamini-Hochberg-Yekutieli** → use for multiple indicators calculated on the same dataset

- The aim is to control the chance of falsely concluding there is a difference when there is none (i.e. making a Type I error)

- Essentially, you require larger differences for statistical significance
Multiple Comparison Adjustment Examples in R

```r
> pvalues<-c(0.003, 0.009, 0.032, 0.078, 0.154)
> adjustedpvaluesBH<-p.adjust(pvalues, method = "BH")
> adjustedpvaluesBY<-p.adjust(pvalues, method = "BY")
> adjustedpvaluesBH
[1] 0.01500000 0.02250000 0.05333333 0.09750000 0.15400000
> adjustedpvaluesBY
[1] 0.0342500 0.0513750 0.1217778 0.2226250 0.3516333
```

In this example, we would **reject the first two null hypothesis** but retain the remaining three with the Benjamini-Hochberg (BH) adjustment (BH).

We would **reject the first null hypothesis** but retain the remaining four with the Benjamini-Hochberg-Yekutieli (BY) adjustment.
Key statistical considerations when inferring difference

- Eyeballing confidence intervals is more conservative than hypothesis testing
- We can make Type I or Type II errors
- Statistical power relates to Type II errors
- Confidence intervals can be used to check power informally
- You need to adjust for multiple comparisons
- Doing so helps to avoid a Type I error
- Multiple comparisons may be made using the same dataset or different datasets → use the appropriate adjustment procedure