Guidance on good practice in statistical interpretation

July 2020
bibliometrics@ucl.ac.uk

The UCL bibliometrics policy sets out some principles for the use of citation metrics in research assessment at UCL. As part of the guidance, this document sets out some general principles for good practice in evaluating and interpreting statistical values.

Expressing a desired property as a single value has a number of benefits, including: quick and easy readability and comparisons; straightforward subsequent calculations and processing (such as totals and averages); and are easy to plot and understand in graphs and figures. For raw data and metrics, such single value metrics are in themselves not inappropriate because they do give the exact values for the defined metric in use; although, these must still be used and interpreted responsibly because they are often used to represent abstract concepts or have different meanings in different contexts (see points 1, 3 and 4 of the policy). However, single value metrics can also commonly be the result of calculating mathematical equations or functions, and assumptions made transitioning among raw, calculated, or aggregated metrics can lead to unintended misuse and misinterpretation. However, in transitioning from many to one number, we lose important information contained in all of the values, and so limits our capacity to understand the nuances and completeness of the metrics. Hence, when responsibly using metrics that result from calculations, we must scrutinise the processes behind their calculation: what data was used, what data is missing; and interpret the metrics under these conditions.

Some very common aggregated metrics include familiar statistics, such as the mean or average. For example, the mean is comprised of all the values in a list of observations divided by the sum of the number of observations. Here, the mean is given as a single value, but in fact represents several values. Again, calculated single value metrics such as the mean have their own meaning and advantages for simplicity and ease of use, and are not inherently problematic. Depending on the question such metrics are entirely appropriate, yet they are still vulnerable to irresponsible use and interpretation. For example, means that are based on small number of observations (i.e., sample size) tend to be more variable and interpretations less reliable. Therefore, it is important to provide or request samples sizes, so one can make an informed decision about whether a mean, for example, is representative of a consistent pattern or the result of a few, possibly by chance, outliers.

Further, when interpreting such aggregated metrics, we must ensure that metrics are applied and interpreted at the correct scale of the subject of investigation. That is, do not apply aggregate level metrics to individual subjects, and vice versa; this is a fundamental logic error termed the ecological fallacy or fallacy of division. In bibliometrics and research evaluations, the most prominent example of this is incorrectly assessing the quality of individual papers based on the impact factor of the journal in which they were published. Here, variation in the number of citations per article within a journal means that assuming the average represents an individual paper would usually misestimate the actual citations of an article. Overestimates are particularly likely because the distribution of citations per article within a journal are usually highly skewed: a few papers extremely well cited, but most with few or no citations. This is why UCL is committed to not using the journal impact factor to assess individual articles.

Often many single value metrics are developed as attempts to represent complex and multidimensional attributes of research or researchers. However, such single metrics tend to fail to represent the desired property because they cannot capture the context-dependencies and multidimensional characteristics of the underlying data, and do not account for variation at the individual level or the circumstances that mediate the value. For example, the h-index attempts to combine publication number with citations to generate a single value indicator of researcher productivity and impact. But (issues of what constitutes “productivity” and “impact” notwithstanding; see policy point 1) the h-index does not provide independent measures of either original metric which themselves are more appropriate in contexts where a more complete understanding of both is required.

Further, the h-index - and metrics in general - should not be used to assess individuals, because they do not account for individual variation due to diversity in research (sub) disciplines and individually variable attributes (e.g., niche research questions, career stage, personal circumstances, and teaching/administrative commitments; see below for more on Author Background). Hence, when assessing researchers for promotion, for example, it is invalid to judge individual researchers with metrics such as the h-index.

At times, considered and appropriately chosen suites of metrics and indicators may better reflect complexity, diversity and comprehensiveness of a subject of investigation. For example, whenever a mean is calculated and quoted, it should be accompanied by at least the number of observations used to calculate it and an estimate of dispersion, variation or distribution of the observations around the mean (such as standard error). In a bibliometric example, total number of citations for a person or organisation is meaningless in itself in most contexts but adding how many publications generated those citations and how many researchers were involved provides a set of metrics with the key information needed to interpret the original citation metric.