How to standardize (if you must)

Article *in* Scientometrics · August 2017

DOI: 10.1007/s11192-017-2495-7

CITATIONS
0

READS
114

3 authors, including:

Marcello D'Agostino
University of Milan
65 PUBLICATIONS **853** CITATIONS
SEE PROFILE

Roberto Ghiselli Ricci
University of Ferrara
31 PUBLICATIONS **365** CITATIONS
SEE PROFILE

Noname manuscript No.
(will be inserted by the editor)

How to Standardize (If You Must)

Marcello D'Agostino · Valentino Dardanoni · Roberto Ghiselli Ricci

Received: date / Accepted: date

Abstract In many situations we are interested in appraising the value of a certain characteristic for a given individual *relative to* the context in which this value is observed. In recent years this problem has become prominent in the evaluation of scientific productivity and impact. A popular approach to such relative valuations consists in using percentile ranks. This is a purely ordinal method that may sometimes lead to counterintuitive appraisals, in that it discards all information about the distance between the raw values within a given context. By contrast, this information is partly preserved by using *standardization*, i.e., by transforming the absolute values in such a way that, within the same context, the distance between the relative values is monotonically related to the distance between the absolute ones. While there are many practically useful alternatives for standardizing a given characteristic across different contexts, the general problem seems to have never been addressed from a theoretical and normative viewpoint. The main aim of this paper is to fill this gap and provide a conceptual framework that allows for this kind of systematic investigation. We then use this framework to prove that, under some rather weak assumptions, the general format of a standardization function can be determined quite sharply.

Marcello D'Agostino
University of Milan
E-mail: marcello.dagostino@unimi.it

Valentino Dardanoni
University of Palermo
E-mail: valentino.dardanoni@unipa.it

Roberto Ghiselli Ricci
University of Ferrara
E-mail: ghsrrt@unife.it

Keywords Standardization · Normalization · z-score · m-score · Location statistics · Dispersion statistics · Citation analysis

1 Introduction

In a wide variety of human affairs we often need to compare data coming from different contexts and make important decisions based on this comparison. For example, a university needs to allocate resources to different departments depending on their scientific productivity over the last five years. In most scientific deparments the accepted measure of productivity is the number of papers published in highly rated international journals. However, mathematicians are known to publish much less than experimental physicists or clinical researchers. Simply comparing the number of publications makes no sense and, yet, a decision has to be made. Similar problems arise when an appraisal of the impact of scientific research is required in terms of some kind of bibliometric analysis.[1] In such situations, decision makers have two options: either they adopt the skeptical view that only homogeneous data can be compared to each other and there is no mathematically well-founded procedure to address the problem of comparing heterogeneous data; or they look for a suitable *standardization* procedure, i.e., a mathematical transformation of the raw data into values that can be meaningfully compared.

According to the first option the answer is the usual one: decisions have to be made case by case, considering all the relevant information that is available and appealing to the discernment ability that belongs to the "tacit dimension" of expert knowledge. However, such a solution may simply not be an option in all situations (e.g., institutional evaluation procedures) in which there is a strong request for *accountability*, which prompts for explicit and public evaluation criteria. In these cases, some kind of standardization may be the only option. Moreover, the less skeptical may find that a suitable standardization procedure can considerably help them making a more informed decision. For this reason, standardization (aka normalization) procedures are playing an increasing role in the evaluation of research especially when it comes to comparing productivity and impact in different fields. There is an active research area that deals with the problem of standardizing citation counts coming from different fields in order to make them comparable. The proposals are based on different criteria. For example some focus on average citation counts (e.g., Abramo et al (2012); Albarrán et al (2011); Li et al (2013); Lundberg (2007); Moed (2010); Radicchi et al (2008); Van Raan et al (2010); Vinkler (2012);

[1] It is not our intention, in this paper, to take a stand on the controversial issue concerning the impact of bibliometric analysis itself, especially when applied to the evaluation of individuals, on the overall quality of research. We only aim at a methodological contribution that is neutral with respect to the different policies that may be adopted to promote the growth of high quality scientific knowledge. However, our analysis may also shed some light on the whole problem and be used to support empirical analysis or simulations of different policies.

Waltman et al (2011); Zhang et al (2014)), others focus on highly cited publications (e.g., Leydesdorff et al (2011); Tijssen et al (2002); Van Leeuwen et al (2003); Waltman and Schreiber (2013)). See Waltman (2016) for a recent and thorough review of citation impact indicators.

However, the use of these procedures is by no means restricted to research evaluation and concerns every human activity in which important decisions need to be made as a result of an appraisal based on data that are not directly comparable. Models based on one or the other standardization procedure are widely used to support decision making in many practical applications, e.g., usability testing (Tullis and Albert, 2013), (Kindlund, 2005), psychological testing (Lezak, 1995), (Kaplan and Saccuzzo, 2013), anthropometry (Wang and Chen, 2012), data mining and cluster analysis (Larose and Larose, 2015), (Milligan and Cooper, 1988), (Stoddard, 1979), composite indicators (Mlachil et al, 2014), (OECD, 2005), to mention just a few. There is therefore a growing need for a conceptual and mathematical characterization of the very notion of standardization that clarifies its scope and limits.

Oddly enough, we are not aware of any systematic effort to analyse the standardization problem from a general theoretical viewpoint. This paper aims to fill this gap and clarify the issue, laying the foundations for further investigation. In this way we hope to shed light on the widespread practice of standardizing raw data: on the one hand, we show to what extent the current practice is justified and what is the general format of a "good" standardization procedure; on the other, we show what are the intrisic limitations of any standardization procedure in terms of the amount of information that it can consistently handle.

In our view, standardizing the raw data means converting them into values that express their relative standing with respect to a group of similar data, i.e., data to which they are deemed to be directly comparable. For example, the scientific productivity of a pure mathematician is relativized to the productivity of the members of his reference context consisting only of pure mathematicians. This is usually done by means of summarizing statistics concerning the reference context. A very popular standardization procedure in some fields (e.g., bibliometrics) is the m-score, which consists in dividing the raw data by their means (Waltman, 2016, Section 6.1). Other quite popular procedures in various fields are: i) the *z-score* that transforms each raw value x by subtracting the mean of the values observed in the reference context and dividing the result by the standard deviation; ii) *max-min normalization*, which consists in subtracting the minimum value observed in the reference context and dividing the result by the range (the difference between the maximum and the minimum values). What are the conceptual grounds of these popular standardization methods? Can they be described as members of a general class of standardization functions characterized on the basis of a small set of plausible axioms? What are the limits on the mathematical form of these functions and on their "resolution", i.e. on the number of summarizing statistics that can be consistently used to partition the raw data into equivalence classes?

There is no claim that standardizing is the ultimate solution to the problem of comparing heterogenous data, nor that such a comparison can always be sensibly made. However, if we *must* standardize, for example because the decision process requires an accountable evaluation procedure, then we show that under some rather weak and natural assumptions, the format of a standardization procedure is determined quite sharply, and so are its intrinsic limitations.

2 The Standardization Problem

In general terms, we are interested in appraising the value x_i that a certain characteristic C takes for a given individual i *relative to* the values $x_1, \ldots, x_n$ that C takes for all the individuals of a certain reference context to which i belongs. For a variety of reasons we may judge that the absolute value x_i is not significant *per se* and what counts is its "relative standing" with respect to the context $x_1, \ldots, x_n$ in which it occurs. Thus, we are interested in the following question:

> How does the relative standing of x_i with respect to its context $x_1, \ldots, x_n$ compare to the relative standing of y_j with respect to its context $y_1, \ldots, y_m$? (Q)

In its general form, the question is somewhat vague and its actual meaning depends on how we interpret the elusive notions of "relative standing" and "context". By a "context" here we simply mean a set of individuals for which we deem that the observed values of a certain characteristic of interest are directly comparable. Clearly, the partition of a population into contexts is the result of a decision that depends on the problem at hand. Here we assume that such a partition has been properly decided.

As for the notion of "relative standing", a natural and popular answer consists in mapping each absolute value to its percentile rank with respect to the values observed in its context. While this approach may work well for certain purposes, it often leads to counterintuitive results. For example, consider two contexts each consisting of 100 values described by the following table:

	1	$\cdots$	9	10	11	12	$\cdots$	100
x	6	$\cdots$	6	6	5	2	$\cdots$	2
y	6	$\cdots$	6	3	2	2	$\cdots$	2

where the values are ordered from highest to lowest. Here the rank of y_{10} is higher than the rank of x_{11} and, yet, some may intuitively judge that the relative standing of x_{11} with respect to $\mathbf{x}$ is higher than the relative standing of y_{10} with respect to $\mathbf{y}$, essentially because the distance between y_9 and y_{10} is much greater than the distance between x_{10} and x_{11}. This intuition may be

captured by transforming the raw values into z-scores and showing that the z-score of x_{11} is higher than the z-score of y_{10}.[2]

Transforming a value into its z-score is a typical example of what is usually called "standardization". In general, by *standardization function* we mean a two argument function S, mapping a raw value x and its context $\mathbf{x}$ to a real number, which is intended to measure the relative standing of x with respect to $\mathbf{x}$.

In our view, a standardization function is one that satisfies the following key property:

> if two values x_i and x_j belong to the same context $\mathbf{x}$, the distance between $S(x_i, \mathbf{x})$ and $S(x_j, \mathbf{x})$ is (strictly) monotonically related to the distance between x_i and x_j. (P)

Clearly, this property is not satisfied by purely ordinal methods. To stick to the above example, the distance between y_9 and y_{10} is much greater than the distance between y_{10} and y_{11}, while the distance between the respective ranks is the same.

How much information do we need about the context $\mathbf{x}$ to be able to standardize a given value x occurring in it? On one extreme, we may think that we need *no* information about $\mathbf{x}$ — say when we feel that comparing absolute values regardless of the context is adequate for our purposes — in which case no standardization is needed. On the other, we may insist that we must know *all* the values of $\mathbf{x}$, in which case any two contexts are essentially different unless they are permutations of each other. In most interesting situations, however, we may be happy with partial information in terms of a suitable finite set of *summarizing statistics* $f_1, \ldots, f_n$ that convey all the information about $\mathbf{x}$ that we deem relevant. The key idea of our approach is that two contexts may be considered *equivalent* whenever they take the same values for all the chosen statistics $f_1, \ldots, f_n$. Equivalent contexts cannot be distinguished in terms of the statistical information that we deem relevant for our purposes and so they should be treated as identical. In other words:

> $S(x, \mathbf{x}) = F(x, f_1(\mathbf{x}), \ldots, f_n(\mathbf{x}))$ for some function F and some fixed set of statistics $\{f_1, \ldots, f_n\}$, (S)

where the choice of the relevant statistics $f_1, \ldots, f_n$ is subjective.

The main aim of this paper is contributing to clarify the question (Q) in terms of standardization functions of the general form (S). More precisely, the abstract standardization problem we address boils down to the following two questions:

Question 1 What sets of statistics can be meaningfully used in (S) to convey all the (subjectively) relevant information about contexts?

[2] Recall that the z-score of w_i with respect to $\mathbf{w} = (w_1, \ldots, w_n)$ is equal to $\frac{w_i - \mu(\mathbf{w})}{\sigma(\mathbf{w})}$, where $\mu(\mathbf{w})$ and $\sigma(\mathbf{w})$ are the mean and the standard deviation of $\mathbf{w}$. In the example the z-scores of y_{10} and x_{11} are respectively 0.5497 and 2.0948.

This involves also considering the possible *limitations* on the amount of statistical information that can be consistently used in (S). In other words: are there constraints on the amount of statistical information that can be consistently used to define a standardization function?

Question 2 Given an appropriate set of statistics, what is the exact functional form of the standardization function?

In the sequel we shall propose a conceptual framework and a set of intuitive properties for a reasonable standardization function that will allow us to answer these questions under rather weak assumptions.

3 The Model

For our purposes we can identify a *context* with the record of values observed for the individuals belonging to it. This amounts to enumerating the individuals in the context and specifying an n-tuple of real numbers where the i-th element corresponds to the value observed for the i-th individual belonging to the context, according to the given enumeration. So, from now on, a context will be formally represented by a subset of the set $\mathbb{R}^n$ of all n-tuples of reals, that we shall often call *vectors*. Let Ω be the set of all finite contexts with at least two elements, formally $\Omega = \bigcup_{n=2}^{\infty} \mathbb{R}^n$. In this paper we consider a *statistic* as any function f mapping each context to a real number, provided that f is *permutation invariant*, i.e. it yields the same value for any two contexts $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{y}$ is simply a permutation of $\mathbf{x}$. For example, given contexts $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} = (0.5, 1.1, 5.6, 0.5)$ and $\mathbf{y} = (1.1, 0.5, 0.5, 5.6)$ it must be the case that $f(\mathbf{x}) = f(\mathbf{y})$. Permutation invariance naturally arises from the widespread assumption of exchangeability of any data vector $(x_1, \ldots, x_n)$. In other terms, the 'labels' identifying the individual data are uninformative, in the sense that the information that the x_i's provide is independent of the order in which they are collected.

As mentioned in the previous section we consider two contexts $\mathbf{x}$ and $\mathbf{y}$ as equivalent with respect to a chosen set of summarizing statistics $f_1, \ldots, f_n$ whenever $f_i(\mathbf{x}) = f_i(\mathbf{y})$ for all $i = 1, \ldots, n$. Intuitively, this means that the two contexts cannot be distinguished from the point of view of the statistical information in which we are interested and, therefore, the data contained in them can be directly compared. For example, if the only statistic information of interest for a given purpose is the mean μ, two contexts $\mathbf{x}$ and $\mathbf{y}$ will be deemed equivalent if $\mu(\mathbf{x}) = \mu(\mathbf{y})$. If we reckon that, besides the mean, also the standard deviation σ and the minimum should be taken into account, then $\mathbf{x}$ and $\mathbf{y}$ will be deemed equivalent if $\mu(\mathbf{x}) = \mu(\mathbf{y})$ *and* $\sigma(\mathbf{x}) = \sigma(\mathbf{y})$ *and* $\min(\mathbf{x}) = \min(\mathbf{y})$. To take a concrete example for this particular choice of statistics, the two contexts $\mathbf{x} = (0, 4, 3, 7)$ and $\mathbf{y} = (7, 4, 0, 3, 6, 1)$ would be equivalent, in that their mean (3.5), their standard deviation (2.5) and their minimun (0) are the same.

More formally, let Φ be an arbitrary finite set of statistics on Ω, let Ω^* be any subset of Ω consisting of the vectors in which we are interested for the purpose of standardization,[3] and let $\sim_\Phi$ be the equivalence relation on Ω^* defined as follows: for all $\mathbf{x}, \mathbf{y} \in \Omega^*$,

$$\mathbf{x} \sim_\Phi \mathbf{y} \text{ if and only if } f(\mathbf{x}) = f(\mathbf{y}) \text{ for all } f \in \Phi. \tag{1}$$

It is immediate to verify that the relation $\sim_\Phi$ is reflexive (for all $\mathbf{x}$, $\mathbf{x} \sim_\Phi \mathbf{x}$), symmetric (for all $\mathbf{x}, \mathbf{y}$, $\mathbf{x} \sim_\Phi \mathbf{y}$ implies $\mathbf{y} \sim_\Phi \mathbf{x}$) and transitive (for all $\mathbf{x}, \mathbf{y}, \mathbf{z}$, $\mathbf{x} \sim_\Phi \mathbf{y}$ and $\mathbf{y} \sim_\Phi \mathbf{z}$ imply $\mathbf{x} \sim_\Phi \mathbf{z}$). Thus, it is an equivalence relation in the mathematical sense. As such, it induces a partition of Ω^* into equivalence classes.

In the following, we shall denote by $[\mathbf{x}]$ the equivalence class of any $\mathbf{x} \in \Omega^*$ associated with the equivalence relation $\sim_\Phi$. In other words, $[\mathbf{x}]$ is the set of all $\mathbf{y}$ such that $\mathbf{y} \sim_\Phi \mathbf{x}$.

Observe that the assumption of permutation invariance for the statistics in Φ avoids the trivial case that $[\mathbf{x}] = \{\mathbf{x}\}$ for any $\mathbf{x} \in \Omega^*$. Given that any $\mathbf{x} \in \Omega^*$ contains at least two elements, then obviously every permutation of $\mathbf{x}$ will also belong to $[\mathbf{x}]$. If we drop this assumption, we don't have the same guarantee, as it occurs, for instance, when Φ is the set of all projections[4] on any Ω^* containing only vectors whose size is bounded above by some $n \in \mathbb{N}$.[5] Anyway, we emphasize the fact that the property of permutation invariance plays absolutely no role in our main result.

Let Φ and Ψ be two arbitrary sets of statistics on Ω: we say that Φ and Ψ are *equivalent* on Ω^* if the following condition holds for all $x, y \in \Omega^*$:

$$\mathbf{x} \sim_\Phi \mathbf{y} \text{ if and only if } \mathbf{x} \sim_\Psi \mathbf{y} \tag{2}$$

Intuitively this means that the choice between the sets of statistics Φ and Ψ does not make any difference as far as the equivalence of contexts in Ω^* is concerned. Clearly, if Φ and Ψ are equivalent on Ω^*, the partition of Ω^* induced by $\sim_\Phi$ is equal to the partition induced by $\sim_\Psi$, that is if $\Omega^*/\sim_\Phi = \Omega^*/\sim_\Psi$. Although this notion of equivalence between sets of statistics can be understood in its abstract sense, obvious cases occur when $\Phi = \{f_1, \ldots, f_n\}$ and $\Psi = \{g \circ f_1, \ldots, g \circ f_n\}$ for some fixed function g (with "$\circ$" denoting the composition of functions). A simplest example occurs when Φ contains only the standard deviation σ and Ψ only the variance σ^2.

We also say that Φ is *redundant* on Ω^* if there exists at least a $\Psi \subset \Phi$ such that Φ and Ψ are equivalent on Ω^*. Simple typical examples are: $\Psi = \{\sigma\}$ and $\Phi = \{\sigma, \sigma^2\}$; $\Psi = \{\mu, \sigma\}$ and $\Phi = \{\mu, \sigma, \frac{\mu}{\sigma}\}$. In general terms, if a set Φ

[3] Depending on the choice of statistics, it may be the case that Ω^* cannot coincide with Ω. For example, if our standardization function is the z-score or the max-min, Ω^* cannot include vectors whose components are all equal, such as (3,3,3,3), for in this case the standard deviation is 0, and so the standardization function would require dividing by 0.

[4] Here by "projection" we mean any mapping from a vector to its value occupying a given position. So, the j-th projection map $proj_j$ is defined as the function mapping each vector $\mathbf{x}$ containing at least j elements on its value x_j.

[5] Recall we assume that Φ is finite.

of statistics is non-redundant, no statistic included in it can be expressed as a function of the others.

Remark 1 Let Φ be an arbitrary set of statistics such that $\{f_1, \ldots, f_k, g\} \subseteq \Phi$. Assume that $g(\mathbf{x}) = F(f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$ for all $\mathbf{x} \in \Omega^*$, where F is any real function defined on $\mathbb{R}^k$. Then, the set Φ is redundant on Ω^*, because it is clearly equivalent to $\Phi \setminus \{g\}$ on Ω^*.

In what follows, all the proofs are relegated to the Appendix.

Lemma 1 *Let Φ be an arbitrary set of statistics on Ω. Let Ω_1 be a subset of Ω such that $\Omega^* \subset \Omega_1$. If Φ is non-redundant on Ω^*, then it is also non-redundant on Ω_1.*

We now turn to our main definition. The underlying idea can be informally explained as follows. Given a set Φ of statistics that, according to our judgment, are sufficient to characterize the equivalence between contexts (with respect to the problem under consideration), our aim is that of defining a meaningful two-argument standardization function of the form (S) (specified in Section 2). This is a function that for any element x_i of a vector $\mathbf{x}$ yields its standardized value $S(x_i, \mathbf{x})$ with respect to $\mathbf{x}$. This function transforms all vectors in Ω^* into their standardized version, in such a way that all the standardized vectors belong to the same "distinguished" equivalence class, so that their values can be directly compared. For this purpose we require that the standardization function satisfies some intuitive constraints specified by A1–A4.

Definition 1 A *standardization set-up* is a quadruple (S, Ω^*, Φ, D) where

(*i*) S is a real function defined on $\mathbb{R} \times \Omega^*$ (called *standardization function*);
(*ii*) Φ is a non-redundant set of statistics on Ω^*;
(*iii*) D is a distinguished equivalence class in the partition of Ω^* induced by the equivalence relation $\sim_\Phi$;

satisfying the following conditions for all $\mathbf{x}, \mathbf{y} \in \Omega^*$

A1) $\bar{S}(\mathbf{x}) \in D$, where $\bar{S}(\mathbf{x})$ is an abbreviation for $(S(x_1, \mathbf{x}), \ldots, S(x_n, \mathbf{x}))$, where n is the size of $\mathbf{x}$.
A2) if $\mathbf{x} \sim_\Phi \mathbf{y}$, then $S(u, \mathbf{x}) \geq S(v, \mathbf{y})$ *if and only if* $u \geq v$;
A3) $S(x_i, \mathbf{x}) = x_i$ *whenever* $\mathbf{x} \in D$;
A4) $|S(u, \mathbf{x}) - S(v, \mathbf{x})| \leq |S(u', \mathbf{x}) - S(v', \mathbf{x})|$ *if and only if* $|u - v| \leq |u' - v'|$ *for all* $u, v, u', v' \in \mathbb{R}$.

In a standardization set-up, S is the standardization function that transforms a raw value x_i occurring in a context $\mathbf{x}$ into a standardized value. The set Φ contains the statistics that we deem sufficient to capture all the relevant information about the contexts. These are used to define the equivalence relation $\sim_\Phi$ that induces a suitable partition on a particular subset Ω^* of Ω. With regard to (iii), D is the class of all *standard contexts*.

From now on, it is implicitly intended that any $f \in \Phi$ is associated with a certain fixed value, for which we will exclusively use the symbol c_f, i.e., the

value taken by f in all standard contexts. For example, if $\Phi = \{f_1, f_2\}$ where f_1 is the mean and f_2 is the standard deviation, typically $c_{f_1} = 0$ and $c_{f_2} = 1$.

Let us now discuss the properties A1–A4. Property A1 states the key property of standardization. It says that, when applied to all the elements of a vector $\mathbf{x}$, a standardization function transforms all contexts into standard ones. Since standards contexts are all equivalent to each other, standardized values can be directly compared. Property A2 implies that when two contexts are equivalent, then S is *strictly increasing* in its first argument and preserves equality.[6] Property A3 says that when the context is already in the distinguished class D of standard contexts, we are happy to leave its values unchanged. Notice that, in conjunction with Property A1, this implies that $\bar{S}(\bar{S}(\mathbf{x})) = \bar{S}(\mathbf{x})$. Property A4 corresponds to the characterizing property (P) of standardization functions discussed in Section 2.

To give an example of a standardization set-up consider the quadruple (S, Ω^*, Φ, D) where $\Phi = \{\mu, \sigma\}$ (i.e, the mean and the standard deviation), $\Omega^* = \{\mathbf{x} \in \Omega : \sigma(\mathbf{x}) \neq 0\}$, $D = \{\mathbf{x} : \mu(\mathbf{x}) = 0, \sigma(\mathbf{x}) = 1\}$ and S is the z-score function. Notice that in this case the set $\Omega^* \subset \Omega$ is the set of all vectors in Ω for which the z-score is defined. Although we are free to choose Ω^* as the set of all contexts on which we are interested for the problem under consideration, the most general case is when it coincides with the set of all contexts for which the standardization function S is defined.

One remarkable consequence of our main result, expressed in Theorem 1 below, is that our definition of standardization set-up dramatically restricts the non-redundant sets Φ of statistics that may occur as third element of the quadruple — i.e., the amount of statistical information that can be consistently used to define a standardization function of the form (S) — to those that *contain at most two elements*. For example, the set $\Phi = \{\mu, \sigma, \min\}$ that we mentioned above to illustrate the equivalence relation $\sim_\Phi$, while being non-redundant, cannot possibly be employed in the definition of a standardization set-up without violating one of the defining properties of S. As we shall see, Theorem 1 also imposes further restrictions on the type of statistics that Φ may contain.

Remark 2 In the following, when we consider an arbitrary standardization set-up (S, Ω^*, Φ, D), by *Lemma 1* we can always assume without loss of generality that Ω^* is *maximal*, in the sense that there does not exist any $\Omega_1 \subseteq \Omega$ such that Ω^* is a proper subset of Ω_1 and (S, Ω_1, Φ, D) is still a standardization set-up.

In other words, Ω^* is maximal when it contains all the vectors in Ω it can possibly contain, provided that the function S is always defined and all the conditions in the definition of a standardization set-up are satisfied.

[6] For technical convenience, we assume that the standardization function is mathematically well-defined even when the first argument is an arbitrary real value that does not belong to the context in the second argument, as is the case with the usual standardization functions. The same assumption is made for Property A4.

For $\alpha > 0$, we say that a statistic f is *positively homogenous of degree* α (briefly, α-positively homogenous) if $f(\lambda\mathbf{x}) = \lambda^\alpha f(\mathbf{x})$ for every $\mathbf{x} \in \Omega$ and every constant $\lambda > 0$. For example, mean, standard deviation, maximum, minimum and range are all positively homogenous of degree 1, for $\mu(\lambda\mathbf{x}) = \lambda\mu(\mathbf{x})$, $\sigma(\lambda\mathbf{x}) = \lambda\sigma(\mathbf{x})$, etc. On the other hand, variance is positively homogenous of degree 2, for $\sigma^2(\lambda\mathbf{x}) = \lambda^2\sigma^2(\mathbf{x})$. When we simply speak of a *positively homogenous* statistic f, we mean that f is α-positively homogenous for some $\alpha > 0$.

Let us say that a vector is *flat* if all its elements are equal and let us write $\mathbf{a}$ to denote the flat vector whose elements are all equal to $a \in \mathbb{R}$. We shall call E the set of all flat vectors. We call any statistic f such that, for all $\mathbf{x} \in \Omega$ and all $\mathbf{a} \in E$, $f(\mathbf{x}+\mathbf{a}) = f(\mathbf{x}) + a$ a *location statistic* . Examples of location statistics are mean, maximum, minimum, median. With a slight terminological abuse[7] we call any statistic f such that $f(\mathbf{x}+\mathbf{a}) = f(\mathbf{x})$ a *dispersion statistic*. Examples are: standard deviation, variance and range.

Remark 3 We emphasize the fact that, unlike the case of dispersion statistics, the only possibility for a location statistic to be positively homogeneous is when the degree α is equal to 1. In fact, let f be a location statistic and fix any $\mathbf{x} \in \Omega, \mathbf{a} \in E$ and $\lambda > 0$: on the one hand, applying in turn the definition of location statistic and α-positive homogeneity, we have that $f(\lambda(\mathbf{x}+\mathbf{a})) = \lambda^\alpha f(\mathbf{x}+\mathbf{a}) = \lambda^\alpha(f(\mathbf{x})+a)$. On the other hand, reversing the order of application of the above properties, we get $f(\lambda(\mathbf{x}+\mathbf{a})) = f((\lambda\mathbf{x}) + (\lambda\mathbf{a})) = f(\lambda\mathbf{x}) + \lambda a = \lambda^\alpha f(\mathbf{x}) + \lambda a$. Therefore, we obtain that $\lambda^\alpha a = \lambda a$, which necessarily implies $\alpha = 1$ by the arbitrariness of a.

Remark 4 If f and g are location statistics, it is immediate to verify that $g - f$ is a dispersion statistic; moreover, according to the previous remark, if f and g are positively homogeneous, so is $g - f$. For example, given the two location statistics max and min, the difference $\max(\mathbf{x}) - \min(\mathbf{x})$ is a dispersion statistic and is positively homogenous (of degree 1).

Now we are ready to state our main result.

Theorem 1 *For every standardization set-up (S, Ω^*, Φ, D) such that Φ contains at least two elements, if all the functions in Φ are location or dispersion statistics and positively homogenous, then:*

T1) *$\Phi = \{f, g\}$ where f is a location statistic and g is either a location statistic or a dispersion statistic;*

T2) *if g is a dispersion statistic, there exists an $\alpha > 0$ such that:*

$$S(u, \mathbf{x}) = \left(\frac{c_g}{g(\mathbf{x})}\right)^{\frac{1}{\alpha}} (u - f(\mathbf{x})) + c_f, \tag{3}$$

[7] Statistics are usually classified into three general classes, that is, *location* statistics (e.g., mean, median, mode, quantiles, minimum and maximum), *dispersion* statistics (e.g., variance, standard deviation, range, interquartile range), and *shape* statistics (e.g., skewness, kurtosis). In our terminology, the class of dispersion statistics includes also that of the shape statistics.

with $\Omega^* = \{\mathbf{x} \in \Omega : g(\mathbf{x}) \neq 0\}$*;*

T3) *if* g *is a location statistic, we have that:*

$$S(u, \mathbf{x}) = \frac{c_g - c_f}{g(\mathbf{x}) - f(\mathbf{x})}(u - f(\mathbf{x})) + c_f, \tag{4}$$

with $\Omega^* = \{\mathbf{x} \in \Omega : g(\mathbf{x}) - f(\mathbf{x}) \neq 0\}$.

While the above theorem shows, somewhat unexpectedly, that the statistical information specified in Φ to establish the equivalence of contexts *cannot consist of more than two statistics*, in the next two corollaries we show what happens in the special case in which Φ contains, respectively, only one dispersion statistic and only one location statistic.

Corollary 1 *Let* (S, Ω^*, Φ, D) *be a standardization set-up such that* $\Phi = \{f\}$*, where* f *is a positively homogenous dispersion statistic. Then, there exist an* $\alpha > 0$ *and a function* $p : \Omega^* \to \mathbb{R}$ *satisfying the conditions*

C1) $p(\mathbf{x}) = p(\mathbf{y})$ *if* $f(\mathbf{x}) = f(\mathbf{y})$*;*
C2) $p(\mathbf{x}) = 0$ *if* $\mathbf{x} \in D$*,*

such that:

$$S(u, \mathbf{x}) = \left(\frac{c_f}{f(\mathbf{x})}\right)^{\frac{1}{\alpha}} u + p(\mathbf{x}), \tag{5}$$

where $\Omega^* \subseteq \{\mathbf{x} \in \Omega : f(\mathbf{x}) \neq 0\}$.

Before stating a corollary analogous to the previous one, we need a preliminary lemma.

Lemma 2 *Let* f *be an arbitrary location statistic. Then, there does not exist any* $r \in \mathbb{R}$ *such that* $f^{-1}(\{r\}) = \{\mathbf{r} = (r, \ldots, r) \in E\}$.

Corollary 2 *Let* (S, Ω^*, Φ, D) *be a standardization set-up such that* $\Phi = \{f\}$*, where* f *is a positively homogenous location statistic. Then, there exists a function* $p : \Omega^* \to \mathbb{R}$ *satisfying* C1 *and the conditions*

C3) $p(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in \Omega^*$*;*
C4) $p(\mathbf{x}) = 1$ *if* $\mathbf{x} \in D$*,*

such that:

$$S(u, \mathbf{x}) = p(\mathbf{x})(u - f(\mathbf{x})) + c_f. \tag{6}$$

Let us now recall the two fundamental questions asked at the end of Section 2, namely: 1) What sets of statistics can be meaningfully used by a standardization function to convey all the (subjectively) relevant information about contexts? 2) Given an appropriate set of statistics, what is the exact functional form of the standardization function?

Theorem 1, together with Corollaries 1 and 2, provide an answer to both our initial questions under the assumption that all the functions in Φ are location or dispersion statistics and positively homogenous. Question 1 is answered

by showing that (i) Φ cannot contain more than two statistics,[8] (ii) at least one of these is a location statistic and at most one is a dispersion statistic. Question 2 is answered by showing that the form of S is completely determined by Eq. (3) or (4).

Note that by Remarks 3 and 4, Eq. (4) is a special case of Eq. (3), so that one can always rearrange a set-up in such a way that Φ contains exactly one location statistic and one dispersion statistic. Besides the z-score function the so-called *robust z-scores*, such as

$$S(u, \mathbf{x}) = \frac{u - \mathrm{med}(\mathbf{x})}{\mathrm{irq}(\mathbf{x})},$$

where med is the median and irq is interquartile range, or

$$S(u, \mathbf{x}) = \frac{u - \mathrm{med}(\mathbf{x})}{\mathrm{mad}(\mathbf{x})},$$

where mad is the median absolute deviation, are also special cases of Eq. (3) (with $\alpha = 1$). Moreover, the equally well-known max-min normalization

$$S(u, \mathbf{x}) = \frac{u - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$$

is a special case of Eq. (4). Notice that, according to our result, the standardization functions of this form make use of a maximal amount of statistical information concerning the contexts in Ω^*, for no further statistical information can be possibly used for standardization purposes.

There exist different examples of standardization functions which make use of a non-maximal amount of statistical information and belong to the models proposed in Corollaries 1 or 2, where $\Phi = \{f\}$ contains only one statistic, such as:

$$S(u, \mathbf{x}) = \frac{u}{\sigma(\mathbf{x})},$$

where f is the dispersion statistic given by the standard deviation, or

$$S(u, \mathbf{x}) = \frac{u}{\max(\mathbf{x}) - \min(\mathbf{x})},$$

where f is the dispersion statistic given by the difference between the maximum and the minimum, or the m-score:

$$S(u, \mathbf{x}) = \frac{u}{\mu(\mathbf{x})}, \tag{7}$$

where f is the location statistic given by the mean, or

$$S(u, \mathbf{x}) = \frac{u}{\max(\mathbf{x})}, \tag{8}$$

[8] Recall that Φ is assumed to be non-redundant.

where f is the location statistic given by the maximum. The first two examples are very special cases of Eq. (5), with $\alpha = c_f = 1$ and $p(\mathbf{x}) \equiv 0$, while Eqs. (7) and (8) are particular cases of Eq. (6), with $c_f = 1$ and $p(\mathbf{x}) = 1/f(\mathbf{x})$. Note that, in the latter case, the maximal Ω^* coincides with the set $\{\mathbf{x} \in \Omega : \max\{\mathbf{x}\} > 0\}$: this point is worth of further comment. In fact, in the literature, it is possible to find standardization functions which correspond to Eq. (8), but such that the respective domains Ω^* are not maximal: this is the case, for example, of the standardization procedure denoted by Z_3 and illustrated in Milligan and Cooper (1988), where the authors specify that the involved variables must be positive or, equivalently in our terminology, that the domain Ω^* must be $\{\mathbf{x} \in \Omega : x_i > 0 \text{ for all } i\}$.

Finally, the rather popular standardization procedure based on ranking (the function Z_7 in Milligan and Cooper (1988)) does not belong to the model investigated in this paper in that it fails to satisfy the characterizing condition (P) discussed in Section 2.

4 An Application

In Zhang et al (2014), the authors analyze the normalization effect of m-score and z-score with respect to raw citations and the relative merits of these two normalization methods. The main approach followed by the authors in order to test the effects of various normalization methods consists in the following steps: (i) put together publications of different fields; (ii) sort them in decreasing order according to their normalized citation counts; (iii) select the global top $t\%$ publications; (iv) distribute them according to their corresponding fields; (v) calculate the percentages of publications falling into the top $t\%$ in each field. Based upon an idea of Radicchi et al (2008), the authors conclude that the most effective normalization method is the one which exhibits the highest uniformity between the global top $t\%$ publications and the top $t\%$ in each field.

We shall illustrate the contribution made by our theoretical analysis to the practical application of such approach for publications belonging to two different fields under an arbitrary standardization procedure.

Let $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be the vectors whose components represent the citations of the m and n papers belonging to two different fields, respectively. We shall assume that the components of both vectors are arranged in decreasing order. After fixing a distinguished set $T = \{t_1, \ldots, t_s\}$ of possible values that t can take, we pick up a specific $t_i \in T$: it is then easy to see that the top $t_i\%$ publications in the first and the second field are given by $\{x_k : k \leq j_m(t_i)\}$ and $\{y_k : k \leq j_n(t_i)\}$, respectively, where $j_p(t_i)$ denotes the integer part of $(t_i\, p)/100$ for any $p \in \mathbb{N}$. Given any standardization set-up (S, Ω^*, Φ, D), owing to A2, we know that the top $t\%$ publications in each single field remain the same regardless of the normalization process. Now, it can be shown that we have a perfect correspondence between the top $t_i\%$ publications

in every field and the global top $t_i\%$ normalized publications if, and only if,

$$\max\{S(x_{j_m(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} \leq \min\{S(x_{j_m(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\}. \quad (9)$$

For instance, suppose that $m = 1500$, $n = 1300$ and $t_i = 1$, i.e. we are interested in the top 1% publications in each field and in the top 1% global normalized ones. Accordingly, $j_m(1) = 15$ and $j_n(1) = 13$, so the top 1% of the publications of the two fields are exactly $\{x_1, \ldots, x_{15}\}$ and $\{y_1, \ldots, y_{13}\}$, respectively. Moreover, according to Eq. (9), our normalization method exhibits the highest uniformity, at least for the percentage of 1%, when

$$\max\{S(x_{16}, \mathbf{x}), S(y_{14}, \mathbf{y})\} \leq \min\{S(x_{15}, \mathbf{x}), S(y_{13}, \mathbf{y})\}.$$

Now, if we analyze in detail Eq. (9), it is immediate to see that we have only the following four cases:

Case (i)

$$\begin{cases} \max\{S(x_{j_m(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} = S(x_{j_m(t_i)+1}, \mathbf{x}); \\ \min\{S(x_{j_m(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\} = S(x_{j_m(t_i)}, \mathbf{x}). \end{cases}$$

Case (ii)

$$\begin{cases} \max\{S(x_{j_m(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} = S(x_{j_m(t_i)+1}, \mathbf{x}); \\ \min\{S(x_{j_m(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\} = S(y_{j_n(t_i)}, \mathbf{y}). \end{cases}$$

Case (iii)

$$\begin{cases} \max\{S(x_{j_m(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} = S(y_{j_n(t_i)+1}, \mathbf{y}); \\ \min\{S(x_{j_m(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\} = S(y_{j_n(t_i)}, \mathbf{y}). \end{cases}$$

Case (iv)

$$\begin{cases} \max\{S(x_{j_m(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} = S(y_{j_n(t_i)+1}, \mathbf{y}); \\ \min\{S(x_{j_m(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\} = S(x_{j_m(t_i)}, \mathbf{x}). \end{cases}$$

The cases (iii) and (iv) are analogous to the cases (i) and (ii), respectively, so it suffices to treat the first two ones. In case (i), Eq. (9) boils down to

$$S(x_{j_m(t_i)+1}, \mathbf{x}) \leq S(x_{j_m(t_i)}, \mathbf{x}),$$

which is assured by property A2, seeing that $x_{j_m(t_i)+1} \leq x_{j_m(t_i)}$ by assumption. Thus, in case (i), as well as case (iii), Eq. (9) is always satisfied. Otherwise, in case (ii), Eq. (9) amounts to

$$S(x_{j_m(t_i)+1}, \mathbf{x}) \leq S(y_{j_n(t_i)}, \mathbf{y}). \quad (10)$$

Suppose now that S satisfies the assumptions of Theorem 1. Then, according to the proof of such theorem (for details see Eq. (16) in Appendix), there exist two *permutation invariant* functions $a, b : \Omega^* \to \mathbb{R}$ such that, for all $u \in \mathbb{R}$:

$$S(u, \mathbf{x}) = a(\mathbf{x})u + b(\mathbf{x}). \quad (11)$$

By applying Eq. (11) yields

$$S(x_{j_m(t_i)+1}, \mathbf{x}) = a(\mathbf{x})x_{j_m(t_i)+1} + b(\mathbf{x}) \tag{12}$$

and

$$S(x_1, \mathbf{x}) = a(\mathbf{x})x_1 + b(\mathbf{x}).$$

The combination of the above formula with Eq. (12) leads to

$$S(x_{j_m(t_i)+1}, \mathbf{x}) = S(x_1, \mathbf{x}) - a(\mathbf{x})(x_1 - x_{j_m(t_i)+1}) \tag{13}$$

and, analogously, it can be shown that

$$S(y_{j_n(t_i)}, \mathbf{y}) = S(y_1, \mathbf{y}) - a(\mathbf{y})(y_1 - y_{j_n(t_i)}). \tag{14}$$

Taking into account Eq. (10), by means of Eq. (13) and Eq. (14) we derive that, in case (ii), Eq. (9) holds true if, and only if,

$$S(x_1, \mathbf{x}) - a(\mathbf{x})(x_1 - x_{j_m(t_i)+1}) \leq S(y_1, \mathbf{y}) - a(\mathbf{y})(y_1 - y_{j_n(t_i)}) \tag{15}$$

and similarly for case (iv). Besides the initial terms $S(x_1, \mathbf{x})$ and $S(y_1, \mathbf{y})$, it is clear that the fulfillment of Eq. (15) is strongly dependent on the crucial function a, which is nothing but the so-called *scaling factor*, whose importance has been emphasized in many papers (see, for instance, Abramo et al (2012)). However, we must not forget that the distribution of the components of the vectors $\mathbf{x}$ and $\mathbf{y}$ can also play a significant role for the satisfaction of Eq. (9). For example, if we compare the m-score and the z-score, whose scaling factors are given by the reciprocals of the mean and the standard deviation, respectively, when the vectors $\mathbf{x}$ and $\mathbf{y}$ are such that $m = n$ and $x_k = y_k + q$ for all $k = 1, \ldots, n$ and for some $q \in \mathbb{N}$, we claim that Eq. (9) is always satisfied for the z-score, but not generally for the m-score. First of all, observe that, as a consequence of the fact that the mean is a location statistic and the standard deviation is a dispersion statistic, we have $\mu(\mathbf{x}) = \mu(\mathbf{y}) + q$ and $\sigma(\mathbf{x}) = \sigma(\mathbf{y})$. Therefore, if we start with the z-score, we obtain that

$$S(x_k, \mathbf{x}) = \frac{x_k - \mu(\mathbf{x})}{\sigma(\mathbf{x})} = \frac{x_k - \mu(\mathbf{y}) - q}{\sigma(\mathbf{y})} = \frac{y_k - \mu(\mathbf{y})}{\sigma(\mathbf{y})} = S(y_k, \mathbf{y})$$

for all $k = 1, \ldots, n$. This immediately implies that, fixing any t_i, both the equalities

$$\max\{S(x_{j_n(t_i)+1}, \mathbf{x}), S(y_{j_n(t_i)+1}, \mathbf{y})\} = S(x_{j_n(t_i)+1}, \mathbf{x}) = S(y_{j_n(t_i)+1}, \mathbf{y})$$

and

$$\min\{S(x_{j_n(t_i)}, \mathbf{x}), S(y_{j_n(t_i)}, \mathbf{y})\} = S(x_{j_n(t_i)}, \mathbf{x}) = S(y_{j_n(t_i)}, \mathbf{y})$$

hold true. In other words, both cases (i) and (iii) occur, hence, as previously shown, Eq. (9) is satisfied for every level t of any arbitrary T. On the contrary, let us prove that with such vectors and the further assumption $0 < y_{j_n(t_i)+1} = y_{j_n(t_i)} < \mu(\mathbf{y})$ the m-score does not verify Eq. (9). We assert that the assumption $0 < y_{j_n(t_i)+1} = y_{j_n(t_i)} < \mu(\mathbf{y})$ implies the occurrence of

case (ii). Indeed, under such hypothesis, an elementary algebraic computation shows that

$$\frac{y_{j_n(t_i)+1}+q}{\mu(\mathbf{y})+q} > \frac{y_{j_n(t_i)+1}}{\mu(\mathbf{y})}.$$

Therefore, the next chain of inequalities directly follows:

$$S(x_{j_n(t_i)+1},\mathbf{x}) = \frac{x_{j_n(t_i)+1}}{\mu(\mathbf{x})} = \frac{y_{j_n(t_i)+1}+q}{\mu(\mathbf{y})+q} > \frac{y_{j_n(t_i)+1}}{\mu(\mathbf{y})} = S(y_{j_n(t_i)+1},\mathbf{y}).$$

In other words, we have shown that

$$\max\{S(x_{j_n(t_i)+1},\mathbf{x}), S(y_{j_n(t_i)+1},\mathbf{y})\} = S(x_{j_n(t_i)+1},\mathbf{x})$$

and, analogously, it can be shown that

$$\min\{S(x_{j_n(t_i)},\mathbf{x}), S(y_{j_n(t_i)},\mathbf{y})\} = S(y_{j_n(t_i)},\mathbf{y}),$$

so proving the assertion. As previously stated, this means that Eq. (9) is fulfilled if, and only if, Eq. (15) is verified. Recalling that $a(\mathbf{z}) = 1/\mu(\mathbf{z})$ for all $\mathbf{z}$ such that $\mu(\mathbf{z}) > 0$, it is not difficult to see that Eq. (15) may be rewritten as

$$\frac{x_{j_n(t_i)+1}}{\mu(\mathbf{x})} \leq \frac{y_{j_n(t_i)}}{\mu(\mathbf{y})}.$$

Since $x_k = y_k + q$ for all k and $\mu(\mathbf{x}) = \mu(\mathbf{y}) + q$, the previous inequality amounts to

$$\frac{y_{j_n(t_i)+1}+q}{\mu(\mathbf{y})+q} \leq \frac{y_{j_n(t_i)}}{\mu(\mathbf{y})},$$

or, equivalently, after a simple algebraic manipulation,

$$y_{j_n(t_i)} - y_{j_n(t_i)+1} \geq q\Big(1 - \frac{y_{j_n(t_i)}}{\mu(\mathbf{y})}\Big),$$

which clearly contradicts our assumption $0 < y_{j_n(t_i)+1} = y_{j_n(t_i)} < \mu(\mathbf{y})$, so definitely proving the claim. As a consequence, at least for the proposed distribution, our conclusion is opposite to the one drawn in Zhang et al (2014), where the authors state that the mean-based method is generally better performing, because it exhibits the highest uniformity between the global top $t\%$ publications and the top $t\%$ in each field. Notice also that this effect, in our theoretical example, is independent of the skewness of the distribution.

5 Conclusions

Theorem 1 is, at the same time, bad news and good news. It is bad news because it shows that standardization is confined by the straight-jacket of set-ups containing no more than two statistics. This provides a strong upper limit on the statistical information that we can consistently use to discriminate between classes of directly comparable data. It is good news because it shows

that a small set of natural requirements on standardization allows us to obtain a quite sharp characterization of its mathematical form.

Our results can be applied in a variety of areas to test whether a proposed standardization procedure complies with them, that is, belongs to one of the classes characterized in Theorem 1, Corollary 1 and Corollary 2. If it does, our notion of standardization set-up and its properties can provide a preliminary mathematical justification for it, restricting the discussion to the most appropriate choice of the location or dispersion statistics to include in Φ, and on whether or not it may be more appropriate to use standardization functions based on one statistic only, or to adopt a more refined version based on two statistics, that is, according to our main theorem, the maximal amount of statistical information that can be managed by a standardization function.

If it does not, our result can still shed light on the problem under investigation, for this implies that one of the properties of a standardization set-up must be violated, and this could be a good reason either for discarding the proposed procedure or for motivating why the violated properties are unsuitable for the problem in question.

Appendix

Proof of Lemma 1

Suppose *ab absurdo* that Φ is redundant on Ω_1. By definition, we know that there exists a proper subset Ψ of Φ such that Φ and Ψ are equivalent on Ω_1. Particularly, this means that $g(\mathbf{x}) = g(\mathbf{y})$ for all $g \in \Phi$ whenever $\mathbf{x}, \mathbf{y} \in \Omega_1$ are such that $f(\mathbf{x}) = f(\mathbf{y})$ for all $f \in \Psi$. However, being Φ non-redundant on Ω^* by assumption, there exist $\mathbf{x}_0, \mathbf{y}_0 \in \Omega^*$ and a $g \in \Phi \setminus \Psi$ such that $f(\mathbf{x}_0) = f(\mathbf{y}_0)$ for all $f \in \Psi$, but $g(\mathbf{x}_0) \neq g(\mathbf{y}_0)$. Since $\mathbf{x}_0, \mathbf{y}_0 \in \Omega_1$, this clearly contradicts the above property, so closing the proof.

Proof of Theorem 1

Let (S, Ω^*, Φ, D) be a standardization set-up such that all the functions in Φ are location or dispersion statistics and positively homogeneous, with $\Phi \supseteq \{f, g\}$.

Observe that it follows from clause A4 in Definition 1 that for all $u, v \in \mathbb{R}$ and all $\mathbf{x} \in \Omega^*$,

$$|S(u+v, \mathbf{x}) - S(u, \mathbf{x})| = |S(u, \mathbf{x}) - S(u-v, \mathbf{x})|.$$

Moreover, $u-(u-v)$ has the same sign as $(u+v)-u$. Applying Property A2, it follows that $S(u+v, \mathbf{x})-S(u, \mathbf{x})$ also has the same sign as $S(u, \mathbf{x})-S(u-v, \mathbf{x})$ and so $S(u+v, \mathbf{x})-S(u, \mathbf{x})$ must be equal to $S(u, \mathbf{x}) - S(u-v, \mathbf{x})$. Notice also that, within a given equivalence class $H \in \Omega^*/\!\sim_{\Phi}$, the standardization function S is independent of its second argument (by A2), since the value of S is preserved under substitutions of the vector in the second argument with an equivalent one. So, let $S_H : \mathbb{R} \to \mathbb{R}$ be the one-argument function defined as follows: $S_H(u) = z$ if and only if $S(u, \mathbf{w}) = z$ for $\mathbf{w} \in H$. Then the above equation can be written as:

$$S_H(u+v) - S_H(u) = S_H(u) - S_H(u-v),$$

for all $u, v \in \mathbb{R}$. By an elementary algebraic manipulation, such property reads as

$$S_H\Big(\frac{u+v}{2}\Big) = \frac{S_H(u) + S_H(v)}{2}.$$

The above equation means that both S_H and $-S_H$ are midconvex on $\mathbb{R}$, and recalling that, being S_H monotone, it is also measurable, as a consequence of a result proved by Blumberg, and independently by Sierpinski (see, for instance, Roberts and Varberg (1973)), it follows that S_H is an affine function. Consequently, there exist real constants a_H and b_H, depending only on H, such that, for all $u \in \mathbb{R}$:

$$S_H(u) = a_H u + b_H.$$

Generalizing the previous argument, we can always associate with any $\mathbf{x} \in \Omega^*$ two real constants, denoted by $a_{[\mathbf{x}]}$ and $b_{[\mathbf{x}]}$, depending only on $[\mathbf{x}]$, such that, for all $u \in \mathbb{R}$:

$$S(u, \mathbf{x}) = a_{[\mathbf{x}]} u + b_{[\mathbf{x}]}. \tag{16}$$

Notice that $a_{[\mathbf{x}]}$ is forced to be positive by A2. We now distinguish the following two cases: (1) Φ contains at least one dispersion statistic and (2) Φ contains no dispersion statistic.

Case 1. $\Phi \supseteq \{f, g\}$ contains a dispersion statistic, say g. Given any $\mathbf{x} \in \Omega^*$, denote by $\underline{b}_{[\mathbf{x}]}$ the vector of E given by $\underline{b}_{[\mathbf{x}]} = (b_{[\mathbf{x}]}, \ldots, b_{[\mathbf{x}]})$. Then, recalling that $\bar{S}(\mathbf{x})$ is an abbreviation for $(S(x_1, \mathbf{x}), \ldots, S(x_n, \mathbf{x}))$ for some $n \in \mathbb{N}$, it follows from (16), the definition of dispersion statistic and the positive homogeneity of g that there exists an $\alpha > 0$ such that:

$$\begin{aligned} g\big(\bar{S}(\mathbf{x})\big) = g\big(a_{[\mathbf{x}]}\mathbf{x} + \underline{b}_{[\mathbf{x}]}\big) &= g(a_{[\mathbf{x}]}\mathbf{x}) \\ &= a^{\alpha}_{[\mathbf{x}]} g(\mathbf{x}). \end{aligned} \tag{17}$$

At the same time, since $\bar{S}(\mathbf{x}) \in D$, owing to A1 we deduce that $g\big(\bar{S}(\mathbf{x})\big) = c_g$, which, combined with Eq. (17), leads to

$$a^{\alpha}_{[\mathbf{x}]} g(\mathbf{x}) = c_g \quad \text{for all } \mathbf{x} \in \Omega^*. \tag{18}$$

We assert that $c_g \neq 0$: otherwise, by Eq. (18), recalling that $a_{[\mathbf{x}]} > 0$ for all $\mathbf{x} \in \Omega^*$, we would obtain that g is identically zero on Ω^*. In this case, it is easy to see that Φ and $\Phi \setminus \{g\}$ are equivalent on Ω^*, so contradicting the requirement that Φ is non-redundant in Definition 1 and proving the assertion. As a straightforward consequence of the assertion and Eq. (18), one finds that $g(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \Omega^*$, showing that Ω^* is a subset of $\{\mathbf{x} \in \Omega : g(\mathbf{x}) \neq 0\}$.

Now, observe that, if Φ contains another dispersion statistic, say g', then, by Eq. (18), it is not difficult to show that there exists an $\alpha' > 0$ such that $g'(\mathbf{x}) = c \cdot (g(\mathbf{x}))^{\alpha'/\alpha}$ for all $\mathbf{x} \in \Omega^*$, where $c = c_{g'} \cdot (c_g)^{-\alpha'/\alpha}$. Consequently, Φ and $\Phi \setminus \{g'\}$ are clearly equivalent on Ω^* against the requirement that Φ is non-redundant. Therefore, Φ can contain at most one dispersion statistic.

Since, by assumption, $\Phi \supseteq \{f, g\}$ and all the functions in Φ are location or dispersion statistics, it follows that f is a location statistic. Repeating the argument illustrated before Eqs. (17) and (18), just replacing the definition of dispersion statistic with the one of location statistic, and recalling Remark 3, we deduce that:

$$\begin{aligned} f\big(\bar{S}(\mathbf{x})\big) = f\big(a_{[\mathbf{x}]}\mathbf{x} + \underline{b}_{[\mathbf{x}]}\big) &= f(a_{[\mathbf{x}]}\mathbf{x}) + b_{[\mathbf{x}]} \\ &= a_{[\mathbf{x}]} f(\mathbf{x}) + b_{[\mathbf{x}]} \\ &= c_f. \end{aligned} \tag{19}$$

Hence, resorting to Eqs. (18) and (19) and after a simple algebraic manipulation, Eq. (16) boils down to Eq. (3). Further, Φ cannot contain any other location statistic, say f'; for, applying again Eq. (19) with f' in place of f, we obtain that $f'(\mathbf{x}) = F(f(\mathbf{x}), g(\mathbf{x}))$, where $F(u, v) = u + cv^{1/\alpha}$, with $c = (c_{f'} - c_f) \cdot c_g^{-1/\alpha}$. Thus, owing to Remark 1, Φ contains a redundant set of statistics given by $\{f, g, f'\}$, against the assumption that Φ is non-redundant. This implies that $\Phi = \{f, g\}$ and S must have the form stated in Eq. (3).

Finally, recalling Remark 2, it is now absolutely clear that Ω^* is maximal if and only if it coincides with the whole set $\{\mathbf{x} \in \Omega : g(\mathbf{x}) \neq 0\}$, so closing the case of the presence of a dispersion statistic in Φ.

Case 2: $\Phi \supseteq \{f, g\}$ contains no dispersion statistic. Then, by assumption, all the statistics in Φ are location statistics, particularly f and g. Now, exploiting Eq. (19), we obtain that

$$b_{[\mathbf{x}]} = c_f - a_{[\mathbf{x}]} f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \Omega^*. \tag{20}$$

Repeating the same argument for g, we have that:

$$a_{[\mathbf{x}]} g(\mathbf{x}) + b_{[\mathbf{x}]} = c_g \quad \text{for all } \mathbf{x} \in \Omega^*. \tag{21}$$

If we insert Eq. (20) into Eq. (21), we get:

$$a_{[\mathbf{x}]}(g(\mathbf{x}) - f(\mathbf{x})) = c_g - c_f \quad \text{for all } \mathbf{x} \in \Omega^*. \tag{22}$$

We claim that $c_g - c_f \neq 0$: otherwise, by Eq. (22), recalling that $a_{[\mathbf{x}]} > 0$ for all $\mathbf{x} \in \Omega^*$, we would derive that $f \equiv g$ on Ω^*, so clearly contradicting the requirement that Φ is non-redundant in Definition 1 and proving the claim. As a straightforward consequence of the claim and Eq. (22), one finds that $g(\mathbf{x}) - f(\mathbf{x}) \neq 0$ for all $\mathbf{x} \in \Omega^*$, showing that Ω^* is a subset of $\{\mathbf{x} \in \Omega : g(\mathbf{x}) - f(\mathbf{x}) \neq 0\}$.

Now, resorting to Eqs. (22) and (20) and after a simple algebraic manipulation, it is easy to check that Eq. (16) boils down to Eq. (4). Further, Φ cannot contain any other location statistic, say f'; for, applying again Eq. (19), with f' in place of f, and Eq. (20), we obtain that $f'(\mathbf{x}) = F(f(\mathbf{x}), g(\mathbf{x}))$, where $F(u, v) = c_1 u + c_2 v$, with

$$c_1 = \frac{c_g - c_{f'}}{c_g - c_f}, \;\; c_2 = \frac{c_{f'} - c_f}{c_g - c_f}.$$

Thus, owing to Remark 1, Φ contains a redundant set of statistics given by $\{f, g, f'\}$, against the assumption that Φ is non-redundant. Therefore, $\Phi = \{f, g\}$ and S must have the form stated in Eq. (4).

Finally, recalling Remark 2, it is now absolutely clear that Ω^* is maximal if and only if it coincides with the whole set $\{\mathbf{x} \in \Omega : g(\mathbf{x}) - f(\mathbf{x}) \neq 0\}$, so closing this case.

This concludes the proof of the theorem.

Proof of Corollary 1

By Case 1 of Theorem 1, after the assignment $p(\mathbf{x}) := \underline{\mathrm{b}}_{[\mathbf{x}]}$, we immediately deduce that S satisfies Eq. (5) and that $\Omega^* \subseteq \{\mathbf{x} \in \Omega : f(\mathbf{x}) \neq 0\}$. Note that C1 is a straightforward consequence of the fact that $\underline{\mathrm{b}}_{[\mathbf{x}]} = \underline{\mathrm{b}}_{[\mathbf{y}]}$ whenever $\mathbf{x} \sim_\Phi \mathbf{y}$, i.e. $f(\mathbf{x}) = f(\mathbf{y})$. Finally, C2 directly stems from condition A3.

Proof of Lemma 2

Suppose *ab absurdo* that such an r exists. Let $\mathbf{a} \in E$: then, by definition of location statistic, we have that $f(\mathbf{r}+\mathbf{a}) = f(\mathbf{r}) + a$ and, at the same time, $f(\mathbf{a}+\mathbf{r}) = f(\mathbf{a}) + r$. Thus, exploiting the assumption $f(\mathbf{r}) = r$, we derive that $f(\mathbf{a}) = a$ for any $\mathbf{a} \in E$. Now, fix any $\mathbf{x} \notin E$: by assumption, we have that $f(\mathbf{x}) = r - h$ for some $h \neq 0$. Then, fixing $\mathbf{h} = (h, \ldots, h) \in E$, we get $f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + h = r$, hence $\mathbf{x} + \mathbf{h} \in f^{-1}(\{r\})$, which is a contradiction, because $\mathbf{x} + \mathbf{h}$ is evidently different from $\mathbf{r}$, since it neither belongs to E.

Proof of Corollary 2

By virtue of Eq. (20), after the assignment $p(\mathbf{x}) := \underline{\mathrm{a}}_{[\mathbf{x}]}$, we immediately deduce that S satisfies Eq. (6). Note that C1 is a straightforward consequence of the fact that $\underline{\mathrm{a}}_{[\mathbf{x}]} = \underline{\mathrm{a}}_{[\mathbf{y}]}$ whenever $\mathbf{x} \sim_\Phi \mathbf{y}$, i.e. $f(\mathbf{x}) = f(\mathbf{y})$. Condition C3 is due to the fact that, as recalled in the proof of Theorem 1, $\underline{\mathrm{a}}_{[\mathbf{x}]}$ is forced to be positive for any $x \in \Omega^*$ by A2. Finally, as direct consequence of the previous lemma, we know that there exists at least a $\mathbf{x} \in D$ such that $x_i \neq c_f$ for some i. Thus, since condition A3 applied to Eq. (6) leads to $p(\mathbf{x})(x_i - c_f) = x_i - c_f$, C4 easily follows.

References

Abramo G, Cicero T, D'Angelo C (2012) How important is choice of the scaling factor in standardizing citations? Journal of Informetrics 6(4):645–654

Albarrán P, Crespo J, Ortuño I, Ruiz-Castillo (2011) The skewness of science in 219 subfields and a number of aggregates. Scientometrics 88(2):385–397

Kaplan R, Saccuzzo D (2013) Psychological Testing. Principles, Applications and Issues. Wadsworth Publishing, Belmont

Kindlund S (2005) A method to standardize usability metrics into a single score. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, pp 401–409

Larose D, Larose C (2015) Data Mining and Predictive Analysis. Wiley & Sons, Hoboken

Leydesdorff L, Bornmann L, Mutz R, Opthof T (2011) Turning the tables on citation analysis one more time: Principles for comparing sets of documents. Journal of the American Society for Information Science and Technology 62:1370–1381

Lezak M (1995) Neuropsychological Assessment. Oxford University Press, Newe York and Oxford

Li Y, Radicchi F, Castellano C, Ruiz-Castillo J (2013) Quantitative evaluation of alternative field normalization procedures. Journal of Informetrics 7:746–755

Lundberg J (2007) Lifting the crown-citation z-score. Journal of Informetrics 1(2):145–154

Milligan G, Cooper M (1988) A study of standardization of variables in cluster analysis. Journal of Classification 5:181–204

Mlachil M, Tapsoba R, Tapsoba S (2014) A quality of growth index for developing countries: a proposal. IMF Working Paper (WP/14/172)

Moed H (2010) Cwts crown indicator measures citation impact of a research group's publication oeuvre. Journal of Informetrics 4:436438

OECD (2005) Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing, Paris

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. In: Proceedings of the National Academy of Sciences of the United States of America, vol 105 (45), pp 17,268–17,272

Roberts A, Varberg D (1973) Convex Functions. Academic Press, New York and London

Stoddard A (1979) Standardization of measures prior to cluster analysis. Biometrics 35:765–773

Tijssen R, Visser M, Van Leeuwen T (2002) Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? Scientometrics 54:381–397

Tullis T, Albert B (2013) Measuring the User Experience. Collecting, Analysing, and Presenting Usability Metrics. Elsevier, Amsterdam

Van Leeuwen T, Visser M, Moed HF, Nederhof TJ, Van Raan AF (2003) The holy grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. Scientometrics 57:257–280

Van Raan A, Van Leeuwen T, Visser N MSand Van Eck, Waltman L (2010) Rivals for the crown: Reply to opthof and leydesdorff. Journal of Informetrics 4:431–435

Vinkler P (2012) The case of scientometricians with the absolute relative impact indicator. Journal of Informetrics 6:254–264

Waltman L (2016) A review of the literature on citation impact indicators. Journal of Informetrics 10:365–391

Waltman L, Schreiber M (2013) On the calculation of percentile-based bibliometric indicators. Journal of the American Society for Information Science and Technology 64:372–379

Waltman L, Van Eck N, Van Leeuwen T, Visser M, Van Raan A (2011) Towards a new crown indicator: Some theoretical considerations. Journal of Informetrics 5:37–47

Wang Y, Chen HJ (2012) Use or percentiles and z-scores in anthropometry. In: Handbook of Anthropometry. Physical Measures of Human Form in Health and Disease, Springer, pp 29–48

Zhang Z, Cheng Y, Liu NC (2014) Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of web of science subject categories.

Scientometrics 101(3):1679–1693