Media diversity and the analysis of qualitative variation

David Deacon & James Stanyer

CENTRE FOR RESEARCH IN COMMUNICATION AND CULTURE, LOUGHBOROUGH UNIVERSITY

January 2019

Abstract

Diversity is recognised as a significant criterion for appraising the democratic performance of media systems. This article begins by considering key conceptual debates that help differentiate types and levels of diversity. It then addresses one of the core methodological challenges in measuring diversity: how do we model statistical variation and difference when many measures of source and content diversity only attain the nominal level of measurement? We identify a range of obscure statistical indices developed in other fields that measure the strength of ‘qualitative variation’. Using original data, we compare the performance of five diversity indices and, on this basis, propose the creation of a more effective diversity average measure (DIVa). The article concludes by outlining innovative strategies for drawing statistical inferences from these measures, using bootstrapping and permutation testing resampling. All statistical procedures are supported by a unique online resource developed with this article.

Key word: Media diversity, diversity indices, qualitative variation, categorical data

Acknowledgements

We wish to thank Orange Gao and Peter Dean of Business Intelligence and Strategy Ltd and Dr Adrian Leguina, School of Social Sciences, Loughborough University for their assistance in developing the conceptual and technical aspects of this paper.

Introduction

Concerns about diversity are at the heart of discussions about the democratic performance of all media platforms (e.g. Entman, 1989: 64, 95; Habermas, 2006: 412, 416; Sunstein, 2018:85). With mainstream media, such debates are fundamentally about plurality, namely, the extent to which these powerful opinion leading organisations are able and/or inclined to engage with and represent the disparate communities, issues and interests that sculpt modern societies. Diversity has become a litmus test for the assessment and regulation of media impartiality. For example, since 2003 the Federal Communications Commission in the USA has seen the promotion of 'viewpoint diversity' as a key rationale for its regulatory activities (McCann, 2013). In the UK the most recent BBC impartiality guidelines state:

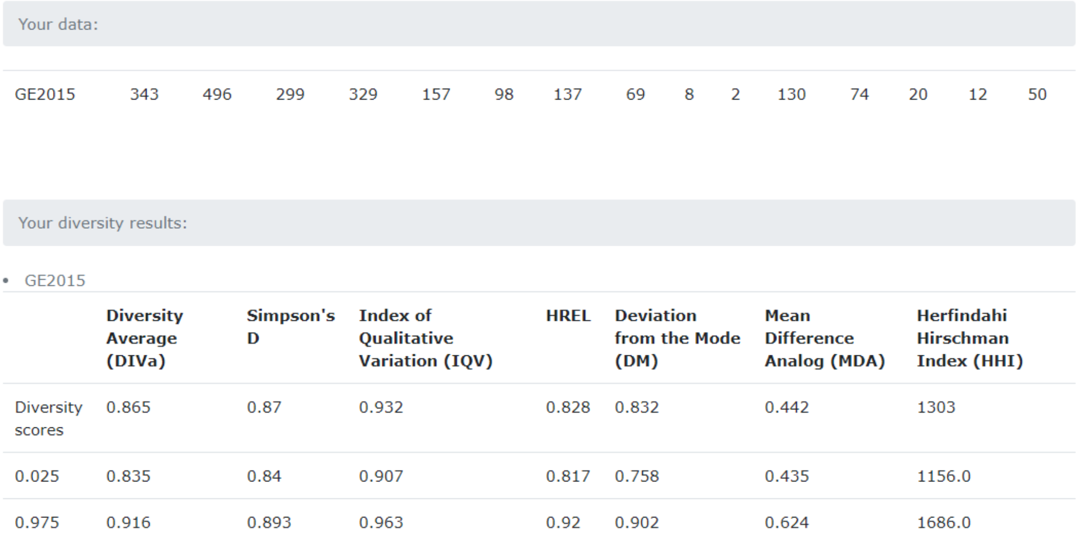
> Across our output as a whole, we must be inclusive, reflecting a breadth and diversity of opinion. We must be fair and open-minded when examining the evidence and weighing material facts. We must give due weight to the many and diverse areas of an argument (BBC, 2018).

This article is mainly focused on the statistical measurement and extrapolation of diversity. We present and assess a range of diversity indices that can be used to measure 'qualitative variation' (i.e. the distribution of values in a variable attaining the nominal level of measurement). These measures remain obscure and poorly understood, both within the field and across the wider statistical literature. In doing so, we propose the creation of a new measure for the analysis of diversity (DIVa: The Diversity Average). We then present two innovative strategies for addressing a major lacuna within the (already limited) literature on diversity measures: how to make statistical inferences on the basis of these descriptive statistical measures. Our analysis is supported throughout by an empirical case study using content analysis data of mainstream news media coverage of three recent 'first order' national electoral events in the United Kingdom.

It is at once an indication of the obscurity of these measures and a barrier to their wider utilisation, that none are supported currently by mainstream statistical software. To help open up their evaluation and use, we have developed a bespoke web resource in conjunction with this article (see figure 1) that allows the automatic calculation of all procedures discussed herein. But any discussion of these technical and methodological matters, needs to begin with consideration of the wider conceptual parameters of the media diversity debate. For what may seem a simple equation (i.e. more diversity = more democracy) is a calculus of greater algebraic complexity. The additional equational components include where one looks for diversity, how one defines it, how it is measured and whether it is possible to conceive of excessive diversity.

Figure 1: DIVA: Diversity Analysis Web Resource

1



Your data:

GE2015	343	496	299	329	157	98	137	69	8	2	130	74	20	12	50

Your diversity results:

- GE2015

	Diversity Average (DIVa)	Simpson's D	Index of Qualitative Variation (IQV)	HREL	Deviation from the Mode (DM)	Mean Difference Analog (MDA)	Herfindahi Hirschman Index (HHI)
Diversity scores	0.865	0.87	0.932	0.828	0.832	0.442	1303
0.025	0.835	0.84	0.907	0.817	0.758	0.435	1156.0
0.975	0.916	0.893	0.963	0.92	0.902	0.624	1686.0

Dimensions of Diversity

Discussions about media diversity connect with all aspects of the circuit of cultural production and consumption. From a supply-side perspective, these include concerns about ownership of media and creative industries, market share and convergence, who works within these industries, and who has the power to regulate their structures and practices. From a demand side, diversity considerations draw attention to questions such as the equality of public access, the needs of citizens, and the predilections of consumers.

Concerns about the diversity of media content and representation intersect with supply and demand side questions, particularly when assessing the impact that different media environments have upon the plurality of public discourse (Roessler, 2008: 476). In discussions about content diversity, attention has focused upon (but is not restricted to) the measurement of *source diversity* (i.e. which individuals and institutions gain greatest prominence in media representations) and *content diversity* (i.e. what issues and frames receive greatest prominence) (Voakes et al, 1996).

These disparate but connected concerns highlight the need to appreciate the different conceptual levels in the analysis of media diversity. McQuail and Van Cuilenburg (1983) distinguish between the *macro level* ('the entire media "system" of a society, regardless of organisational or channel differences' (ibid: 151)), the *meso level* ('a sector within a media system.' (ibid.)) and the *micro level* ('the individual media organisation or outlet - a particular newspaper, periodical or television channel' (ibid.). In a comprehensive review of the literature, Roessler (2008) also uses these terminological distinctions but in a somewhat different way. Whereas McQuail and Van Cuilenburg's typology mainly relates to content diversity, Roessler draws a useful distinction between 'diversity' ('the variety or breadth of media content available to media consumers' (p. 467) and 'diversification' ('the supply side of media firms' ibid.). In his analysis, the macro-level attends to 'the media system and its overall *structure* (2008: 465, emphasis added), the meso-level ('single media outlets in a given media system' (ibid.)) and the micro level (the coverage 'of issues and protagonists' (ibid.))[i].

These different levels of analysis in turn raise the question: at what level should assessments about the content diversity be made? Should the focus be on diversity *within* particular media outlets (variously referred to as 'internal', 'vertical' or 'intra-media' diversity) or on cumulative diversity *across* media sectors or systems (referred to as 'external', 'horizontal' or 'inter media' diversity) (for discussions see McQuail and Van Cuilenburg, 1983: 151-152 and Van Cuilenburg, 2000: 28-9)? There are tensions between these perspectives as '[t]he more intra diverse content packages are, the less inter diverse they can be – and vice versa' (Van Cuilenburg, 2000: 28). Understanding these alternate stances is also valuable for making conceptual distinctions within and between national media systems. For example, UK national newspapers are renowned for their political partisanship and party alignment, whereas public service broadcasters are required to demonstrate due impartiality and balance in their coverage of, and relations with, the political sphere. The justification for the former is found in appeals to 'external diversity' (i.e. citizens can choose a newspaper that most suits their political palate in a competitive market place), whereas the rationale for the latter relates to the importance of securing 'internal diversity' (because of the centrality of broadcasting channels as providers of entertainment and information and, in the case of the BBC, its public funding)[ii]. According to Sheppard (2007: 10), the stance of the UK press contrasts with that in the US, as the latter has become less partisan over time, marking a shift historically from external to internal diversity.

One of the challenges of research in media diversity is to understand the nature and conditions of these intersecting levels and concepts. For example, Voakes et al (1996) used an empirical case study to assert that 'the common assumption that source diversity begets content diversity is fallacious. They sometimes accompany each other, but they appear to vary independently of each other' (p.591). Other studies also challenge assumptions that diversity of ownership guarantees greater plurality of content. For instance, George (2002) and Berry and Waldfogel (2001) (both cited in Roessler, 2008) separately found that greater concentration in ownership increased rather than reduced content diversity and formats – the explanation being that media owners sometimes seek to differentiate content provision across outlets in their portfolio to increase overall consolidation of market share (Roessler, 2008: 478)

Diversity can sometimes be too much of a good thing. Just as a non-diverse mainstream media can narrow and constrain the parameters of public understanding, so excessive diversity can create a disabling diffuseness in public discourse (Gitlin, 1998). This in turn begs the question as to what normative principles should be applied when appraising media diversity. One approach orientates to the concept of 'reflective diversity' and measures the goodness-of-fit between media representations and known population distributions and preferences. The other is 'open diversity' which is more concerned with the opinion-leading potential of the media in social terms and their responsibility to extend and pluralise public debate, regardless of underlying social configurations. Van Cuilenberg notes the tensions between reflective and open diversity: 'Media fully reflecting social preferences inevitably ill perform at openness to a greater variety of different perspectives, social positions and conditions, whereas perfect media openness harms majority positions in favour of minority perspectives, beliefs, attitudes and conditions' (2002: 30).

Measure for Measure: statistical description of diversity

A shared aspect of all these conceptual debates is a fundamental concern with questions of accumulation, scale and patterning across media content, contexts and temporalities. These matters invite questions of measurement and particularly their statistical presentation and summation. In this respect there is a challenge, as many things that are measured to assess media diversity only attain the nominal/ categorical level of measurement. For example: how concentrated are ownership patterns within and between media markets? How do gender and ethnicity intersect with professional status and financial reward in particular creative industries? Which frames and whose voices tend to be prioritised and marginalised within a chosen issue domain?

This reliance on categorical measures precludes many of the standard measures of variance applied to data that attain ordinal, interval and ratio levels. It also means that many of the statistical analyses of differences and trends in diversity are reliant principally on cross-tabulated comparisons (e.g. Deacon et al., 2017, Wahl-Jorgensen et. al, 2017). Such tabulations have undeniable analytical value: it is important to demonstrate where observed differences or confluences occur. But their use and interpretability rapidly degrade the more data points and categories that are introduced and as the analysis extends into multi-variate dimensions. This is particularly regrettable for the analysis of diversity as it limits opportunities to model the relationship between the different conceptualisations outlined earlier (e.g. how the strength in correlation between source and content diversity varies with differential orientation to internal or external diversity norms?).

It so happens there exist a range of statistical indices that measure the strength of 'qualitative variation' (i.e. statistical dispersion across nominal variables) and are potentially of great value in addressing these methodological challenges. These indices have been developed across a variety of disciplines, including botany, economics, sociology and information science, but what they all share is obscurity. Writing several decades ago, the statistician Allen Wilcox noted that 'the discussion of the measurement of variation with nominal-scale data is usually conspicuous by its absence' (1973: 325) and we can attest, having trawled dozens of contemporary statistical textbooks from a range of disciplines in preparing this article, that this remains the case. They are referenced infrequently and their calculation is not supported by major statistical software such as SPSS. As noted below, there are instances where these measures have been used in the analysis of media performance and diversity, but these occasions are vanishingly small and we contend it is timely for researchers within the field to consider their use, application and limitations more widely.

In developing our analysis, two earlier interventions have proven particularly valuable in identifying the main indices. The first is *Indices of Qualitative Variation and Political Measurement* authored by the aforementioned Wilcox (1973). The second is *The Conceptualisation and Measurement of Diversity* written by Daniel McDonald and John Dimmock (2003). Fifteen different indices are described across both reviews, but, as a further indicator of the inchoate nature of this statistical field, there is almost no overlap in the measures appraised (although, as is discussed below, HREL and Shannon's H are closely linked).

Table 1: Diversity Indices in Wilcox (1973) and McDonald and Dimmock (2003)

Wilcox (1973)	**McDonald and Dimmock (2003)**
Deviation from the Mode (DM)	Simpson's D
Mean Difference Analog (MDA)	Simpson's D_z
Average Deviation Analog (ADA)	Kvalseth's OD
HREL	Jung's H
Variance Analog (VA)	Shannon's H
Kaiser's B	Gleason's D
	Hall & Tideman's H
	Fager's S
	Fager's NM

At first sight, both reviews appear to exclude two further indices of qualitative variation that have been used relatively widely. The Herfindahl-Hirschman Index (HHI) was developed in economics to measure market concentration and has been applied in several studies of media diversity (e.g. Entman, 2006, Powers and Benson, 2014). The Index of Qualitative Variation (IQV) has achieved some prominence as a diversity measure in sociological research (e.g. Agresti and Agresti, 1978, Marsden, 1987). The apparent exclusion of each from these reviews has different explanations. McDonald and Dimmock (2003: 67-68) explain that HHI is mathematically equivalent to Simpson's D but not as simple to interpret (NB the version of Simpson's D included in this article produces a diversity measure between 0-1 whereas HHI scores can range from nearly 0 to 10,000). The absence of the IQV measure is due to Wilcox labelling it the 'Variance Analog' in his overview. Terminological inconsistency of this kind is a consistent (and confusing) feature of the use of these statistical measures, as are differences in the calculation of some indices with the same name. For example, Shannon's H, first developed by the information theorist Claud Shannon in the 1940s (Shannon, 1948), is also referred to as the 'Shannon entropy measure', the 'Shannon Weaver index' and the 'Shannon Weiner index' (for an explanation of this confusion, see Spellerberg & Fedor, 2003). It is also often found that H is then standardised equitability[iii], to produce a final statistic between zero and 1. This is variously referred to as HREL or the Shannon Equitability Index. Similar confusions exist with Simpson's D, which is sometimes referred to as the Gini-Simpson's index, and depending on its mode of calculation, can generate figures between zero or one, or from 1 and above.

Given the variety of indices available, the assessment we provide in this article will focus on the statistical indices deemed by Wilcox, McDonald, Dimmock and other authors to be most effective, stable and easy to use (e.g. see also Tan and Weaver, 2013: 778, Teachman, 1980: 344). These are: Simpson's *D*, *HREL* (the standardised version of Shannon's H), the Index of Qualitative Variation (*IQV*) (aka Variance Analog in Wilcox's typology), the Mean Difference Analog (*MDA*) and the Deviation from the Mode (*DM*). All of these measures meet Wilcox's (1973) requirements that: (1) variation is measured as between one and zero; (2) When the distribution of all of the observations are the same, variation is zero; and (3) When all of the observations are equally different, the variation is one. They also pass the dual-concept requirement identified by McDonald and Dimmock (2003). This is that the measures take account of: (1) the range of discrete categorisations within a nominal distribution, and (2) the assignment of elements to those categories. With all indices, nominal distributions can be entered as observed values or percentage distributions. Additionally, categories that contain zero observations must be included in the calculation as their emptiness says something about the overall diversity of a distribution.

Table 2 outlines the statistical equations we used to calculate each measure (with notation standardised across the various equations). As mentioned earlier, we have developed a web-based resource that automates these calculations.

Table 2: Equations for diversity indices

Simpsons D	$1 - \frac{\sum f(f-1)}{N(N-1)}$
Index of Qualitative Variation	$\frac{K(100^2 - \sum Pct^2)}{100^2(K-1)}$
HREL	$-\sum_{i=1}^{k} \frac{f_i}{N} \log_2 \frac{f_i}{N}$

Mean Difference Analog	$1 - \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\lvert f_i - f_j\rvert}{N(K-1)}$
Deviation from the Mode	$1 - \frac{\sum_{i=1}^{K}(f_m - f_i)}{N(K-1)}$

Notes: Notation has been standardised across the equations

f = the number of cases in a category.

N = the number of cases in a sample.

K = the number of categories in a distribution.

Pct= the percentage a category represents of the total number of cases.

*f*m = the number of cases in the modal category.

*f*i = the number of cases in the *i*th category.

The case study

To provide insight into the use and performance of these selected measures we used data from media content analyses of three recent UK electoral events: national TV and newspaper reporting of the 2015 United Kingdom General Election (GE2015), the 2016 UK European Union Membership Referendum[iv] (REF2016) and the 2017 UK General Election (GE2017). All studies were conducted by the Centre for Research in Communication and Culture, Loughborough University. Each study was a discrete analysis, which means these are comparisons of cross sectional studies rather than a formal longitudinal design. Nevertheless, the studies used repeated measures, identical sampling procedures[v], near identical coding teams and were each subject to inter-coder reliability testing.

To assess the performance of the selected indices, we compared their measurement of 'source diversity' in the coverage of these three electoral events (GE2015, REF2016 and GE2017). Our measure of source diversity was *the number of occasions that party political representatives were directly quoted in news and feature items related to each campaign sample*.

The first issue we examined was whether and to what extent different category groupings of the same data sets changed diversity scores - i.e. could we identify any categorisation effects? As noted, one of McDonald and Dimmock's (2003) 'dual concept' requirements for evaluating a diversity measure is that the calculation takes account of the range of discrete categorisations within a nominal distribution, thereby opening up the potential for the comparison of diversity scores for distributions with different numbers of categories. However, if diversity scores appear to be affected by the range of categorisations, such practices need to be questioned.

To analyse this aspect, we categorised source diversity distributions for the three campaigns in three ways (see Table 3). Variable A differentiated between sources from the two main parties and placed 'all other parties' in a single category. Variable B elaborated the 'all other parties' category to produce a distribution with eight party categories. Variable C extended this further, making a distinction between the appearance of 'party leaders' and 'all other party sources' for the 7 main UK political parties and a remaining category for 'all other party sources', producing 15 categories. Each of these categorisations can be conceived as appropriate for measuring different facets of the debate about the diversity of party political representation in the mainstream media. Variable A provides a measure of the extent and parity of bi-partisanship in campaign coverage. Variable B provides a more sensitive measure of the range and depth of multi-party representation. Variable C permits the analysis of intra-party diversity in coverage, in particular, the extent to which coverage has become more or less dominated by party leaders. The categorisations and the data distributions for each are set out in Table 3. Table 4 provides the statistical summary for source diversity in each campaign and categorisation, using the five selected diversity indices.

In broad terms, the diversity measures for each categorisation of the three campaigns tell a consistent story: all indicate that coverage of the 2015 UK general election had the highest source diversity, followed by the 2017 UK General Election and then the 2016 EU Referendum. More detailed analysis of these results reveals variation both within and between diversity measures. This variation is not consistent, but there are some tendencies. The IQV results were either the highest or joint highest in 7 of the 9 cases, and the second highest in the remainder. At the other end, MDA scores were the lowest in 7 of the 9 cases, and second lowest in the remainder. The remaining three measures fluctuated more inconsistently in positioning. D scores were second highest in 5 cases, but the lowest in 2 of the remainder, DM scores were never higher than third in the rankings, HREL results varied most widely in terms of rankings.

It is worth noting in passing the implications this last point holds for a tendency we have noted within the economics literature around the use of the Herfindahl-Hirschman Index (HHI) in the measurement of market competition. As noted earlier, the HHI is the direct mathematical equivalent of Simpsons' D, but produces results from close to zero to 10,000 in which the lower a score is, the greater the diversity of a distribution. US government agencies currently apply an abstracted standard that judge markets with an HHI below 1500 as 'unconcentrated', with an HHI between 1500 and 2500 as 'moderately concentrated' and an HHI above 2500 above HHI as 'Highly concentrated' (see US Department of Justice, 2018). If we were to use the GE2015 distribution as a proxy for a measure of concentration within an imaginary market, the Variable B categorisation would produce an HHI score of 2529, and the Variable C, an HHI of 1303. In other words, applying the abstracted US government agency criteria, the Variable B categorisation would indicate a 'highly concentrated' market and the Variable C categorisation an 'unconcentrated' one.

Table 3: Categorisations for party sources: GE2015, REF2016, GE2017

Variable A				**Variable C**			
	GE2015	**REF2016**	**GE2017**		**GE2015**	**REF2016**	**GE2017**
	%	%	%		%	%	%
Conservative	37.7	64.7	43.3	Cons Leader	15.4	12.5	18.4
Labour	28.2	22.0	35.3	Cons other	22.3	52.2	24.9
Other Party	34.0	13.4	21.4	Labour Leader	13.4	4.1	16.0
(Number)	(2224)	(1588)	(1867)	Labour other	14.8	17.9	19.3
				Lib Dem Leader	7.1	0.4	5.0
Variable B				Lib Dem other	4.4	1.3	3.5
	GE2015	**REF2016**	**GE2017**	SNP Leader	6.2	0.9	2.9
Conservative	37.7	64.7	43.3	SNP other	3.1	0.8	1.0
Labour	28.2	22.0	35.3	UKIP Leader	5.8	6.8	2.4
Lib Dem	11.5	1.6	8.5	UKIP other	3.3	2.2	1.6
SNP	9.3	1.6	3.9	Green Leader	0.9	0.1	0.7
UKIP	9.2	9.0	4.0	Green other	0.5	0.1	0.1
Green	1.4	0.2	0.7	PC Leader	0.4	0.0	1.1
Plaid Cymru	0.4	0.0	1.3	PC other	0.1	0.0	0.3
Other party	2.2	0.9	2.9	Other party	2.2	0.9	2.9
(Number)	(2224)	(1588)	(1867)	(Number)	(2224)	(1588)	(1867)

Table 4: Macro-Diversity Scores for Media Coverage of GE2015, REF2016 and GE2017

	Campaign	HREL	D	IQV	MDA	DM
Variable A	**GE2015**	0.993	0.662	0.993	0.907	0.944
	REF2016	0.804	0.516	0.773	0.487	0.53
(3 categories)	**GE2017**	0.965	0.643	0.963	0.781	0.851
Variable B	**GE2015**	0.761	0.747	0.854	0.396	0.712
(8 categories)	**REF2016**	0.524	0.525	0.600	0.163	0.404
	GE2017	0.670	0.677	0.773	0.308	0.648
Variable C	**GE2015**	0.828	0.87	0.932	0.442	0.832
	REF2016	0.590	0.673	0.721	0.177	0.512
(15 categories)	**GE2017**	0.756	0.835	0.894	0.353	0.805

Where does this recognition of the variance across diversity indices leave us in terms of selecting a measure? We discern two approaches within the limited literature on diversity indices. The first might be labelled 'the first past the post' approach and involves asserting that one measure is the 'best' overall measure (e.g. Tan and Weaver, 2013). The second, to extend the horse racing parlance, can be termed the 'horses for courses' approach and claims that some measures are better for addressing specific questions related to diversity and some for others (e.g. Teachman, 1980: 344). In our view, the first approach presents a question that is difficult to resolve, whereas, the second poses a question that is difficult for non-specialists to understand. A neat and simple solution, in our view, would be to develop an integrated measure that averages a composite of selected measures. Straightforward averaging might seem crude, but it is a technique that is commonly used in areas such as economic forecasting and is recognised as improving forecast accuracy (see Bates and Granger, 1965, Clemen, 1989).

For the constituent elements of this measure we recommend using the four diversity measures that performed most consistently according to the previous tests (HREL, D, IQV and DM). We propose naming this composite measure the Diversity Average (DIVa) and have included a facility for its automated computation on the web resource developed in conjunction with this article.

Table 5 outlines the DIVa scores for all media in the three campaigns using the Variable A, B and C categorisations. Once again, we calculated the average of all permutations of the absolute differences, for each election and found DIVa had a lower averaged variation than all single measures (0.097). This suggests that the combination of measures in its calculation ameliorates the categorisation effects noted with individual measures.

Table 5: Diversity Averaging: DIVa Scores by Campaign

	2015	2016	2017
DIVa Variable A	0.898	0.656	0.855
DIVa Variable B	0.768	0.513	0.692
DIVa Variable C	0.865	0.624	0.822

From Description to Inference: the need for innovation

Thus far, our analysis has identified statistical variations in the source diversity indices for the three campaigns, with GE2015 revealing the greatest source diversity in media coverage, followed by GE2017 and then REF2016 media campaigns. These measures are descriptive statistics and as such are limited in the extent to which they can be extrapolated and compared. The variations may reveal something important in differences across the three media campaigns or they may just be a product of the random fluctuation that occurs in any form of sampling.

Statistical inference permits two core tasks: estimating ‘true’ population parameters from descriptive measures and hypothesis testing (Deacon et al., 2007). Wilcox notes that it is necessary to gain knowledge of the sampling distributions of diversity indices ‘if one hopes or intends to apply procedures of statistical inference’ (1973: 340). In this respect, the already limited literature on the measurement of qualitative diversity is nearly silent. Neither McDonald and Dimmock (2003) nor Wilcox (1973) provide guidance on statistical testing and most studies that use diversity indices do not include significance tests. On the rare occasions such tests appear, it is difficult to identify the procedures that have been used (e.g. Humprecht and Esser, 2017) or their application requires advanced understanding and operationalisation of statistical formulae (e.g. Agresti and Agresti, 1978: 211-229).

Elsewhere, there are occasional references to the potential utility of the Mann-Whitney U test in connection with diversity indices, but no worked examples are provided (e.g. Fowler et al., 2002, p.102).

In this section, we propose the use of ‘bootstrapping’ and ‘permutation testing’ techniques to address the twin challenges of drawing population estimates and hypothesis testing from observed diversity scores. Bootstrapping can be used to calculate confidence intervals for population estimates (i.e. the range in which the ‘true’ population value likely to lie on the basis of sampled observations). Permutation testing can be used for hypothesis testing (i.e. calculating whether observed differences in descriptive measures are statistically significant). Both approaches have long prehistories (e.g. Fisher, 1935), but it is only with the accessibility to powerful computer technology that their widespread application has become realistic (Mooney, 1996). Both deploy resampling techniques to generate the necessary sampling distributions needed for inferential statistical work. Part of the distinction between the two lies in the nature of the resampling undertaken (see below).

Bootstrapping

The name ‘bootstrapping’ is taken from the phrase ‘pulling oneself by one’s own bootstraps’ and calculates population distributions by using the information you have available: i.e., your sample data (hence the allusion to self-elevation in the title). The process involves generating large numbers of random resamples (with replacement) from this data set. When we talk of sampling ‘with replacement’ we are describing a process whereby once a value is selected for inclusion in a resample, it is returned to the selection process thereby permitting the possibility that it might again be resampled randomly (and therefore potentially appear more than once in the resample).

Bootstrapping for the diversity measures discussed in this article involved generating a 1000 resamples with replacement of the samples for GE2015, REF2016 and GE2017. We then calculated all selected diversity indices for each resample and ranked the distribution of each from highest to lowest. This produced a bootstrap distribution that is used as our sample distribution. To calculate the 95 percent confidence interval with 1000 resamples ($p<0.05$), we identified the 26th lowest ranked value and the 975th highest ranked value for each diversity measure. Table 6 shows the CIs for DIVa, D, HREL, IQV and DM. Its details show, for example, that from the observed D diversity score for the sampled media for GE2015 of 0.870, there is a 95% probability that the ‘true’ diversity during this campaign ranged between 0.840 and 0.893[vi].

Table 6: 95% Confidence Intervals estimated for GE2015, REF2016 and GE2017 Variable C Source Diversity Scores

		GE2015: 95 % Confidence	
	Diversity score	*-0.025*	*-0.975*
DIVa	0.865	0.834	0.915
D	0.87	0.84	0.893
HREL	0.828	0.816	0.919
IQV	0.932	0.908	0.963
DM	0.832	0.753	0.902

REF2016: 95% Confidence Interval

		-0.025	- 0.975
DIVa	0.624	0.572	0.746
D	0.673	0.587	0.744
HREL	0.59	0.595	0.758
IQV	0.721	0.663	0.835
DM	0.512	0.434	0.653

GE2017: Confidence Interval

		-0.025	- 0.975
DIVa	0.822	0.792	0.878
D	0.835	0.8	0.864
HREL	0.756	0.748	0.866
IQV	0.894	0.87	0.935
DM	0.805	0.726	0.869

Permutation testing

Permutation testing shares similarities to bootstrapping, in that it involves the estimation of population distributions by multiple resampling of observed values. The principal differences are that it is used to compare differences between two variables (hence its role in hypothesis testing) and the resampling is conducted *without replacement* (i.e., when a value is resampled it cannot be included in subsequent selections). The procedure starts by considering 'the value of the statistic actually observed in the study' (Hesterberg et al, 2007: 16.42). In our example, this would be comparing the observed differences in two-way comparisons of QV scores for source diversity in media coverage of different campaigns and it is computed by calculating the absolute value between the two diversity scores (e.g. in GE2015, Simpson's D=0.870 and in REF2016=0.673 = an observed difference of 0.197). We then construct a null hypothesis – i.e. that the observed difference between the two scores is likely to be the product of chance and cannot be deemed statistically significant. The next step requires the creation of a sampling distribution that 'this statistic would have if the effect were *not* present in the population' (ibid.: 16.40), and then comparing the observed statistic to this distribution. If the value sits centrally in the distribution there is a high chance it occurred by chance, but the further away from the centre it sits, the greater the probability 'that something other than chance is operating' (ibid).

The sampling distribution described above is known as the permutation distribution and assumes that the null hypothesis is true (i.e. 'that the two groups' code counts are identically distributed, or that the grouping variable does not influence the outcome' [Collingridge, 2013: 89]). To construct this distribution, the observed distributions in two samples are shuffled into a pooled sample from which two new samples are created that replicate the original sample structures. These twinned samples are created by randomly selecting sample units that are not then returned to the pooled data (i.e. resampled without replacement). The measure under assessment (e.g. each of the selected diversity scores) is then calculated for both resamples and the value of the second resample is subtracted from the first, to produce a positive or negative value. This process is repeated multiple times to produce the permutation distribution (once again, 1000 resamples should be seen as a lower limit). The observed difference is then mapped onto the permutation distribution to gain a *p* value. This is calculated as the proportion of resampled values that give a result as high as the observed difference being tested. For example, we have noted that the difference in the Simpsons' D diversity score for GE 2015 and REF2016 was 0.197. Resampling results found 13 of the 1000 resamples produced a value the same or higher, which means the estimated *P*-value is 0.013 (13/1000).

There is one important caveat to be borne in mind when using this method. Significance testing will not operate for the comparison of variables with 3 or fewer categories. This is because if there are less than 20 possible permutations with two independent variables, $p \leq 0.05$ can never be attained (see Ludbrook and Dudley, 1998: 130).

Table 7: Hypothesis testing: Variable C comparisons

Comparison of GE2015 and REF2016	**GE2015**	**REF2016**	**Absolute difference in diversity scores**	***P***	**sig p<0.05?**
DIVa	0.865	0.624	0.241	0.01	Yes
D	0.870	0.673	0.197	0.01	Yes
HREL	0.828	0.590	0.238	0.02	Yes
IQV	0.932	0.721	0.211	0.01	Yes
DM	0.832	0.512	0.320	0.01	Yes
Comparison of GE2017 and REF2016	**GE2017**	**REF2016**	**Absolute difference in diversity scores**	***P***	**sig p<0.05?**
DIVa	0.822	0.624	0.198	0.05	Yes
D	0.835	0.673	0.162	0.05	Yes
HREL	0.756	0.590	0.166	0.07	No
IQV	0.894	0.721	0.173	0.05	Yes
DM	0.805	0.512	0.293	0.05	Yes
Comparison of GE2015 and GE2017	**GE2015**	**GE2017**	**Absolute difference in diversity scores**	**P**	**sig p<0.05?**
DIVa	0.865	0.822	0.043	0.221	No
D	0.870	0.835	0.035	0.181	No
HREL	0.828	0.756	0.072	0.159	No
IQV	0.932	0.894	0.037	0.186	No
DM	0.832	0.805	0.028	0.36	No

Table 7 shows the results of two-way permutation testing for each of the three media campaigns. The results show the degree of statistical confidence we can have that these observed sample differences between paired media campaigns (i.e. GE2015 and

REF2016; GE2017 and REF2016; GE2015 and GE2017) are probabilistically indicative of actual differences in the wider population. The findings reveal that the observed differences in source diversity between GE2015 and REF2016 for all indices are statistically significant (using $p \leq 0.05$ to determine statistical significance), whereas the differences between GE2015 and GE2017 are not. The results also show that differences between GE2017 and REF2016 are statistically significant for D, IQV and DM, but not for HREL (P=0.07). This raises a further question about the use of individual diversity measures, as this example suggests that the determination of statistical significance can be affected by the choice of measure. This adds strength to the case for developing a composite measure like DIVa, as it offers a way of resolving those instances where significance tests for different measures provide discrepant conclusions.

Conclusion

This article has analysed the key conceptual debates about media diversity and identified a major methodological challenge: how can statistical variance be modelled across these different conceptual frameworks when so many of the core measures only attain the nominal level of measurement?

We have shown that a valuable way forward is to make greater use of various indices of qualitative variation that have been developed across a wide range of disciplines. These measures are not part of the statistical mainstream and this obscurity is compounded by the confusion and inconsistencies surrounding their labelling. Identical measures are sometimes given different names (e.g. Shannon's H/ Shannon Weaver's H/ Shannon Weiner's H). On other occasions, measures with the same name are differently calculated to produce results that require different interpretation (see the standardised and non-standardised ways of calculating Simpsons' D). Elsewhere, certain measures produce different statistical outcomes that are nonetheless functionally identical (Simpson's D and the Herfindahl-Hirschman Index). One of our aims in this paper has been to identify and resolve these confusions.

We have also assessed the performance of five diversity indices: Simpsons' D, the Index of Qualitative Variation (IQV), HREL, the Mean Difference Analog (MDA) and the Deviation from the Mode (DM). The performance of each was demonstrated through original analysis of 'source diversity' in mainstream media coverage of three recent UK electoral campaigns (the 2015 General Election, the 2016 EU Referendum and the 2017 General Election). This analysis shows that some measures tend to give higher standardised diversity scores than others when applied to the same data and that statistical outcomes are affected by the number of categories used to measure qualitative variation. There are important implications to both of these points: diversity scores always need to be interpreted contextually and when comparisons are made between data distributions there is a need to ensure they share the same categorisation structures. The statistical analysis also identified concerns about the reliability of the Mean Difference Analog, to the extent that we question its value as a measure of qualitative variation.

This evaluation leads us to recommend the development of a new averaged measure of the most stable diversity indices. This kind of approach is commonly found in financial forecasting and is recognised to improve forecast accuracy. We propose labelling this new composite measure DIVa: the Diversity Average and show that it helps avoid, on the one hand, the contentiousness of claiming that one measure is superior to others (the 'first past the post' argument) and, on the other, the complexities of determining which measures are best for which scenarios (the 'horses for courses' argument). The measure also addresses, at least partially, the two problems noted above, in that it flattens out fluctuations across particular measures, and reduces categorisation effects.

We have also developed innovative methodological strategies to address one of the most neglected aspects of the literature on diversity indices: how to make statistical inferences from these descriptive measures? To estimate confidence intervals, we outline how 'bootstrapping' resampling methods can be used to construct sampling distributions on the basis of observed values. For hypothesis testing, we show how 'permutation testing' resampling permits the creation of a null hypothesis distribution that then allows the calculation of the probability that an observed difference between two diversity scores occurred by chance. The latter exercise also reveals a further value to the new DIVa measure, as it helps to arbitrate those occasions when certain diversity indices suggest there is a significant difference between two data distributions and others indices, do not.

All of the challenges outlined above demand computational assistance (assessing diversity measures, comparing their performance, creating composite measures and conducting inferential statistical tests on their basis). The absence of an accessible resource to perform these tasks has been a major hindrance to the wider application and evaluation of these measures and developing this article has required us to produce our own computational procedures to facilitate all of these tasks. The fruits of this activity are now offered freely to the field as an online resource to facilitate further investigation into the value of the quantification of qualitative variation.

References

Agresti, A. and Agresti, B. (1978). Statistical analysis of qualitative variation. *Sociological Methodology*, 9, 204 -237.

Bates, J. and Granger, C. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), 451-468.

BBC (2018) *Editorial Guidelines, Section 4: Impartiality* (Accessed 17 April 2018 http://www.bbc.co.uk/editorialguidelines/guidelines/impartiality/breadth-diversity-opinion).

Berry, S. T., & Waldfogel, J. (2001). Do mergers increase product variety? Evidence from radio broadcasting. *The Quarterly Journal of Economics, 116,* 1009–1025. https://doi.org/10.1162/00335530152466296.

Clemen, R. (1989). Combining forecasts: a review and bibliography with discussion. *International Journal of Forecasting*, 5, 559-608.

Deacon, D., Pickering, M., Golding, P., and Murdock, G. (2007) *Researching Communications: A Practical Guide to Media and Cultural Analysis*, London: Bloomsbury.

Deacon, D. and Wring, D. (2017). One party, two issues: UK news media reporting of the EU Referendum. In J. Mair, T. Clark T, R. Snoddy and R. Tait (Eds.) *Brexit, Trump and the Media*, (pp.36-44). Bury St Edmonds: Abramis.

Deacon, D., Downey, J., Smith, D., Stanyer, J. and Wring, D. (2017). The Media Campaign: The Issues and Personalities Who Defined the Election. In J. Mair, T. Clark T, R. Snoddy and R. Tait (Eds.) *Brexit, Trump and the Media*, (pp.367-371). Bury St Edmonds: Abramis.

Entman, R. (2006). Punctuating the homogeneity of institutionalized news: abusing prisoners at Abu Ghraib versus killing civilians at Fallujah. *Political Communication*, 23 (2), 215–224. https://doi.org/10.1080/10584600600629844.

Entman, R. (1989) *Democracy without Citizens: Media and the Decay of American Politics*, New York: Oxford University Press.

Fisher, R.A. (1935) *The Design of Experiments*, 3rd Edition, London: Oliver & Boyd.

Fowler, J., Cohen, L., Jarvis, P. (2002) *Practical Statistics for Field Biology* (Second edition), London: Wiley.

George, L. M. (2002). Ownership concentration and product variety in daily newspaper markets. In *Communications policy and information technology: Promises, problems, prospects,* L. F. Cranor & S. Greenstein (Eds.), Cambridge, MA: The MIT Press. (pp. 235–251).

Gitlin, T. (1998). Public sphere or public sphericules?. In *Media, Ritual, Identity,* T. Liebes & J. Curran, eds., London: Routledge. (pp.168-175).

Habermas, J. (2006). Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research. *Communication Theory*, Vol 16, 411–426. https://doi.org/10.1111/j.1468-2885.2006.00280.x

Hesterberg, T., Moore, D., Monaghan, S., Clipson, A., Epstein, R. and Craig, B. (2007). Bootstrap Methods and Permutation Tests. In *Introduction to the Practice of Statistics*, Moore, D., McCabe, G., Craig, B., eds) (pp.16.1-16.58), 6th edition, New York: W. H. Freeman.

Humprecht, E. and Esser, F. (2017). Diversity in online news. *Journalism Studies*, 19(12), 1825-1849. https://doi.org/10.1080/1461670X.2017.1308229.

Ludbrook, J. and Dudley, H. (1998) Why permutation tests are superior to *t* and *F* tests in biomedical research, *The American Statistician*, 52:2, 127-132, DOI: 10.1080/00031305.1998.10480551

Marsden, P. (1987). Core discussion networks of Americans. *American Sociological Review*, 52, 122-131.

McCann, K. (2013)**.** *The diversity policy model and assessment of the policy***.** *SAGE Open*, 3(2): 1-12. https://doi.org/10.1177/2158244013492780.

McDonald, D. and Dimmock, J. (2003). The conceptualisation and measurement of diversity. *Communication Research*, 30(1), 60-79. https://doi.org/10.1177/0093650202239026.

McQuail*, D.* and J. Van Cuilenburg (*1983*). Diversity as a media policy goal: a strategy for evaluative research and a Netherlands case study. *Gazette* 31(3), 145-162.

Mooney, C. (1996). Bootstrap statistical inference: examples and evaluations for political science. *American Journal of Political Science*, 40(2), 570-602.

Powers, M. and Benson, R. (2014). Is the Internet homogenizing or diversifying the news? external pluralism in the U.S., Danish, and French press. *The International Journal of Press/Politics*, Vol. 19(2), 246–265. https://doi.org/10.1177/1940161213519680.

Roessler, P. *(*2008*).* Media content diversity*:* Conceptual issues and future directions for communications research. In: *Beck*, *CS* (*ed*.) *Communication Yearbook 31 (*pp 464–520). *New York*: Lawrence Erlbaum.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Sheppard, S. (2007). *The Partisan Press: A History of Media Bias in the United States*. Jefferson: McFarland & Co.

Simpson, E.H. (1949). Measurement of diversity. *Nature*, 163, p688.

Spellerberg, I. & Fedor, P. (2003). A tribute to Claude Shannon (1916 – 2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Weiner' Index. *Global Ecology and Biogeography,* 12, 177-179. https://doi.org/10.1046/j.1466-822X.2003.00015.x

Sunstein, C. (2018) *#Republic: Divided Democracy in the Age of Social Media*, Princeton: Princeton University Press.

Teachman, J. (1980). Analysis of population diversity: measures of qualitative variation. *Sociological Methods and Research*, 8(3),341-362. https://doi.org/10.1177/004912418000800305

Tan, Y. & Weaver, D. (2013). Agenda diversity and agenda setting: from 1956 to 2004. *Journalism Studies*, 14(6), 773-789. https://doi.org/10.1080/1461670X.2012.748516

Van Cuilenburg, J. (2000). Media diversity, competition and concentration: concepts and theories. In De Bens, E (ed.) *Media Between Culture and Commerce* (pp. 25-54). Changing Media – Changing Europe Series (Volume 4). Bristol: Intellect,.

Voakes, P. Kapfer, J. Kurpius, D. and Shano-yeon Chern, D. (1996). Diversity in news: a conceptual and methodological framework. *Journalism & Mass Communication Quarterly*, 73(3), 582-593.

Wahl-Jorgensen, K., Berry, M., Garcia-Blanco, I., Bennett, L., Cable, J. (2017). Rethinking balance and impartiality in journalism? How the BBC attempted and failed to change the paradigm. *Journalism*, 18(7), 781:800. https://doi.org/10.1177/1464884916648094

Wilcox, A. (1973). Indices of qualitative variation and political measurement. *Western Political Quarterly*; 26(2): 325-343.

[i] See also Van Cuilenburg's (2000) 4-level distinction between (1) individual content units, (2) content bundles, (3) medium type and (4) the communication system 'as a whole'.

[ii] Having noted this, the BBC impartiality guidelines do make some concession to the concept of external diversity, accepting that in certain circumstances impartiality requirements can be met 'in series and over time' (see http://www.bbc.co.uk/editorialguidelines/guidelines/impartiality/impartiality-series-time#mr. Accessed 30 May 2018).

[iii] This is done by dividing the observed diversity (H) by the maximum potential diversity that could be achieved (H_{max}).

[iv] The referendum vote took place on 23 June 2016 in the United Kingdom (UK) to measure public support for the UK either remaining in or leaving the European Union (EU). Fifty two percent of voters supported the UK's departure from the EU.

[v] For each electoral event, we sampled the last twenty full week days of coverage (excluding polling day) from the following news outlets: (TV News) Channel 5, 6.30pm; Channel 4, 7pm; Sky News, 8-8.30; BBC1, 10pm; ITV, 10pm; (Newspapers) the Guardian, The Times, Daily Telegraph, Financial Times, Daily Mail, Daily Express, Daily Mirror, The Sun and the Star. For TV news, we scrutinised entire programmes for election related coverage. For newspapers, we monitored the front pages, the first two pages of the domestic news section, the first two pages of any specialist election campaign section and the page containing and facing papers' leader editorials.

[vi] There will be an inevitable, small degree of fluctuation in CIs for diversity scores when processes are repeated. This is due to the randomisation involved in the multiple re-samplings.