The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Evaluating Wikipedia as a self-learning resource for statistics: You know they'll use it

Peter K. Dunn, Margaret Marshman & Robert McDougall

To cite this article: Peter K. Dunn, Margaret Marshman & Robert McDougall (2017): Evaluating Wikipedia as a self-learning resource for statistics: You know they'll use it, The American Statistician, DOI: 10.1080/00031305.2017.1392360

To link to this article: https://doi.org/10.1080/00031305.2017.1392360

Accepted author version posted online: 30 Oct 2017.

Evaluating Wikipedia as a self-learning resource for statistics: You know they'll use it

Peter K. Dunn, Margaret Marshman, and Robert McDougall

Peter K. Dunn (E-mail: *pdunn2@usc.edu.au*) is Associate Professor in Biostatistics, Margaret Marshman (E-mail: *mmarshma@usc.edu.au*) is Senior Lecturer in Mathematics Education, and Robert McDougall (E-mail: *rmcdouga@usc.edu.au*) is Lecturer in Mathematics, Faculty of Science, Health, Education and Engineering, University of the Sunshine Coast, Sippy Downs, 4558, Queensland, Australia.

Abstract

The role of Wikipedia for learning has been debated because it does not conform to the usual standards. Despite this, people use it, due to the ubiquity of Wikipedia entries in the outcomes from popular search engines. It is important for academic disciplines, including statistics, to ensure they are correctly represented in a medium where anyone can assume the role of discipline expert. In this context, we first develop a tool for evaluating Wikipedia articles for topics with a procedural component. Then, using this tool, five Wikipedia articles on basic statistical concepts are critiqued from the point of view of a self-learner: "arithmetic mean", "standard deviation", "standard error", "confidence interval" and "histogram". We find that the articles, in general, are poor, and some articles contain inaccuracies. We propose that Wikipedia be actively discouraged for self-learning (using, for example, a classroom activity) except to give a brief overview; that in more formal learning environments, teachers be explicit about not using Wikipedia as a learning resource for course content; and, because Wikipedia is used regardless of considered advice or the organizational protocols in place, teachers move away from minimal contact with Wikipedia towards more constructive engagement.

Key words

statistical language; accuracy; display; statistical definitions

1. INTRODUCTION

Wikipedia is a "free-access, free-content Internet encyclopaedia" (Wikipedia 2016) launched in 2001. Wikipedia now appears in over 200 languages, and the English language version (the largest) boasts over 5 million articles. Wikipedia was the fifth most popular website in the world on 3 April 2017 (Alexa Internet 2017).

One distinctive feature of Wikipedia is the article editing policy: most articles can be edited by anyone with Internet access. This policy has led many to question the accuracy of Wikipedia articles, since editors are not restricted to discipline experts. Wikipedia articles were once driven by experts (Kittur et al. 2007), but more recently there has been "a dramatic shift in workload to the 'common' user" (p. 1). Nonetheless, one review of 42 scientific articles (Giles 2005) found that Wikipedia's accuracy is commensurate with that of *Encyclopaedia Britannica*, though aspects of this study were questioned, including the narrowness of the articles examined (Yasseri et al. 2012).
Wikipedia has been criticized for showing systemic bias (Distaso 2012), containing falsehoods (Black 2010), being subject to spin and manipulation on controversial topics (Petrilli 2008), being edited with inappropriate intent, being edited for the sake of disruption (Search Engine People, 2015), or being edited by researchers wanting to see how long intentionally-inserted falsehoods take to be corrected (Soylu 2009; Magnus 2009; Halavais as cited in Read (2006); Tynan (2008)).

In response, those managing Wikipedia protocols imposed three requirements on authors: articles need to (i) be written from a neutral point-of-view; (ii) contain verifiable claims; and (iii) contain no original research (Wikipedia 2016). These criteria are a good fit for well-established and less controversial disciplines like mathematics and statistics, and as there is a substantial knowledgeable user base from which editors may emerge, it may be that Wikipedia articles on basic mathematical and statistical topics are of high quality, and of broad general utility.

Many people struggle with basic statistical and mathematical concepts (Chance et al. 2004; Fidler and Cumming 2005; Vaux 2012; Dunn et al. 2016b), including, but not

restricted to, occasional users of statistics, workers in non-quantitative disciplines who occasionally need to use mathematical and statistical concepts, those who may not be familiar with the English language, and school or university students learning the topic (Waller 2011). In all these cases, Wikipedia articles may be accessed by "self-learners" to learn more about the topic (Soylu 2009). For these groups, Wikipedia is often a useful information source, either directly or as one of the more recognizable (and often first) web pages returned by an Internet search (Julien et al. 2009), even though better resources (even online) usually exist. Despite this, the role and purpose of Wikipedia is not to act primarily as a comprehensive learning resource but more as a quick reference tool. While there are many opinions about the value of Wikipedia articles, part of our motivation is to provide criteria and an evidence-base for framing and informing opinions on the use of Wikipedia for self-learners (and other learners). In this paper, we evaluate some basic statistical articles to measure the quality of the information contained therein.

We have two aims: to develop a tool for evaluating Wikipedia articles (and similar resources) that describe concepts with a procedural dimension; and to use this tool to evaluate five Wikipedia articles for some basic statistical concepts. After background material, we proceed through a series of steps (Print 1993) to generate an evaluation of relevant articles allowing design characteristics to emerge, conclusions to be drawn and recommendations made. In doing so, we acknowledge Wikipedia's dynamic content of varying stability, and hence this paper represents a snapshot at a point in time.

2. BACKGROUND

As noted above, many users of Wikipedia articles are learners, if not formally as students. For this reason, exploring the role of Wikipedia for self-learning is reasonable.

The role of Wikipedia as a learning resource has been debated, since it contradicts the traditional academic view. For example, Wikipedia lacks academic authority and provides no assurance of expert peer review (Soylu 2009; Viégas et al. 2004). Magnus (2009) lists five criteria to assess claims from sources, and decides that Wikipedia "frustrates all of them" (p. 74). Furthermore, Wikipedia articles are not necessarily

written by experts (Kittur et al. 2007), and collective editing can be problematic (Lanier 2006; Soylu 2009; Read 2006). This may lead to incompleteness (a study of thousands of political articles found that "errors of omission are extremely frequent" (Brown 2011, p. 339)) and inconsistencies (articles usually have many different authors, who may not all have the same goal, style, notation, knowledge, area of application, or intent). Wikipedia articles also may be fluid, and information appearing in an article one day may not be there the next. Wikipedia also has the potential, or the perception of the potential, for sabotage: Anyone can edit most articles, even for malicious means, as noted above.

Despite these potential problems, or the perception of these problems, using Wikipedia can have advantages. In general, Wikipedia articles are up-to-date and can change rapidly when circumstances change, unlike printed material, and cutting edge developments can be announced simply without the time lag that occurs for professional literature and gains wide exposure quickly (P. Sturgess, personal communication, July 2, 2015). Articles are also easily accessed online (which may appeal to the online generation), and Wikipedia is well known. Wikipedia articles may also serve as an entry point to topics and provide a first point of contact with the literature (Waters 2007). Furthermore, using many different authors, combined with the neutral-point-of-view policy, helps opposing points of view to have a voice.

Schools and universities are divided in their response (Read 2006) with recommendations like "you should be extremely cautious about using Wikipedia" (President and Fellows of Harvard College 2016) and "students should not use *Wikipedia* as a substantive source in research papers" (Magnus 2009, p. 88). In most (if not all) cases, these responses are based on anecdotal evidence; in this paper, we hope to provide more robust evidence. Whatever the opinions held by academics and teachers, many self-learners refer to Wikipedia when doing research (Selwyn and Gorard 2016; Head and Eisenberg 2010; Waller 2011): "it would be naïve to tell students that they ought never to use *Wikipedia*… They will use it, and so will we" (Magnus 2009, p. 88). Some authors, therefore, do not ask *whether* people should use Wikipedia or not, but *how*.

3. ON THE ROLE OF WIKIPEDIA AS A STATISTICS LEARNING RESOURCE

Many statistical concepts such as "standard deviation" are well-established and not controversial (though definitions may vary slightly), so some of the reasons for discouraging the use of Wikipedia may not apply in subjects such as statistics. As a result, this paper asks: what value do Wikipedia articles have for those wishing to learn about topics from introductory statistics, and hence what role should Wikipedia play?

The self-learner has many sources of information from which to choose. Introductory statistics books and textbooks are plentiful; however, ready access may be problematic and a library visit not warranted, and these resources are probably more detailed than needed. In addition, textbooks are prepared for specific audiences (see Dunn et al. 2016 a), whereas Wikipedia is a one-size-fits-all model. Many excellent websites also exist, but self-learners are not likely to find them.

We acknowledge that the purpose of Wikipedia is that "each entry in Wikipedia must be about a topic that is encyclopedic and is not a dictionary entry or dictionary-like" (Wikipedia 2016), meaning its purpose is not necessarily for independent learning. However, since many self-learners use Wikipedia in this way. We evaluate some articles through this lens as a pragmatic legitimate interest, not as a criticism.

Previously, Wikipedia has been evaluated for accuracy (Giles 2005), readability (Thomas Jefferson University 2010), trustworthiness (Zeng et al. 2006), inter-article links (Adafre and de Rijke 2005; Schönhofen 2006), semantic relatedness (Strube and Ponzetto 2006), and collection evolution (Almeida et al. 007). While important, the scope of this article demands other issues be evaluated that are specific for our purpose, so we begin by developing a tool for evaluating articles.

4. METHODS

4.1 Framework

We have adapted the process described by Print (1993) as a framework of curriculum evaluation, since we are considering the use of Wikipedia as a self-learning resource. This framework (p. 211, 213) includes six steps:

1. Presage: Clarify what is to be evaluated about each article.
2. Task specification: Establish which articles to evaluate.
3. Evaluation of design and methods: Determine how the information will be collected to enable the evaluation.
4. Data collection.
5. Data analysis.
6. Conclude and report.

We now evaluate five statistical Wikipedia articles using this framework.

4.2 Presage: Step 1

Wikipedia has its own process for evaluation of article importance and quality (Wikipedia: Version 1.0 Editorial Team 2016). For example, the Statistics WikiProject (WikiProject Statistics 2016) includes 3508 entries (as of 5 August 2016), ranked according to their *importance* and *quality*. However, the criteria used to assign these rankings in Wikipedia are not necessarily aligned with our needs.

An article's *importance* is generally rated as Top, High, Mid or Low (WikiProject Mathematics 2016). The entry for Confidence Intervals is not among those assessed as Top Importance, showing that an article's importance defined by Wikipedia does not necessarily align with some of the basic concepts our target users may be considering.

The *quality* of Wikipedia articles may be assigned as one of FA (Featured Article); A; GA (Good Article); B+; B; C; Start; or Stub (described at WikiProject Mathematics 2016). Some articles (such as Disambiguation pages) have other classifications. A Featured Article (Wikipedia: Featured Article Criterion 2017) refers to a small number of articles (approximately 0.1%; Wikipedia: Featured Articles 2017) that meet broad criteria, including content-based criteria. (Only two statistical articles have been Feature Articles (Category: FA-Class Statistics Articles 2017): "Actuary", and "Confirmation bias".) The other criteria are non-content based, showing that the Wikipedia quality criteria are mostly editorially based. Interestingly, these criteria would be useful as applied to a text-based resource, and none of the criteria refer to the opportunities afforded by the online environment, such as videos, animated GIFs and audio files. Specifically, none of these criteria are related to how suitable the entries are for learning about a topic.

For this work, rating a Wikipedia article's usefulness considered the relevant GAISE criteria (Aliaga et al. 2010); curriculum-based criteria established by Print (1993, p. 214), namely interest, authenticity, appropriateness, organization and balance, and technical quality; the online scaffolding and learning style criteria of McDougall et al. (2003) adapted to the Wikipedia environment; and Upchurch's (2011) support for the CRAAP test developed by the Meriam Library (2010) to evaluate information, including currency using relevance, authority, accuracy and purpose.

On their own, none of the sets of criteria above are suited to the purpose of supporting the self-learning of basic statistical concepts for general users, so we have used the most relevant aspects to build an empirical composite frame:

A. how **accurate** is the content? (The content refers to definitions; interpretation; notation; usage; examples. The accuracy includes errors, ambiguities, omissions and inconsistencies.)

C. how effective are **conceptual explanations** in the article? (Explanation of concepts leads to procedures; explanation of concepts beyond the definitions; explanation of what's behind the procedure.)
P. how effective are **procedural explanations** in the article? (Procedure is accurately explained; examples used to explain procedure; explanation of procedure.)
D. how effective is the **display** or presentation of the material? (Clear; accessible; coherent and well-paced; organized; logical; interesting; context; readability; density of formulae; use of diagrams, videos, animations etc. for illustration; complexity of mathematics.)

These criteria include three concerning content, and one concerning presentation. Explanations have been separated into both procedural (knowing how to do it) and conceptual (knowing why it is done).

In practice, the most useful way to review an article is to match criteria to the different components of articles: the whole article, the text appearing before the Contents (here called the Preamble), and the first paragraph of the Preamble. (For example, the Preamble and its first paragraph are easily accessed on mobile devices with small screens, and in any case may be all that is read.) The authors used these criteria to evaluated articles collectively, using a three-point ordinal scale: 1 corresponds to "Don't recommend: Not suitable for self-learning"; 2 corresponds to "Suitable for self-learning when supplemented with authoritative resources"; and 3 corresponds to "Recommended for self-learning". To clarify why a rating of 3 is not applied, a label of A, C, P and/or D (called a "limitation"), corresponding to the above criteria, is given to indicate perceived shortfalls.

We also recorded two text-based qualities, used previously (Elia 2009) to compare Wikipedia articles with *Encyclopaedia Britannica* articles. We assessed readability (Gunning Fog score, the level at which articles are written): 6 is easy to read (upper primary/elementary school) while a score of 13 is more difficult (about first-year university level). We also assessed the lexical density (To et al. 2013; Linnarud 1976),

the amount of information that a selection of text contains (not a measure of readability or level of complexity), expressed as a percentage; a high lexical density percentage means that the text is rich in information. We measured both using the online resource Textalyser (http://textalyser.net/). The Textalyser results should not be over-interpreted. For the whole article, we provided the URL to Textalyser and so images and navigational elements are included, which influence the score. For the Preamble, we cut-and-paste the text (including images but no navigation elements) into Textalyser. For this reason, the results are best used for comparing similar components of articles across the articles chosen for this investigation. All such indices are guidelines only, since using any automated method has limitations (Hartley 2016; Begeny and Greene 2014).

In addition to evaluating individual articles, we evaluate the consistency in notation and language across the suite of articles. In total, each article receives three ratings (one each for the article, the Preamble, and the first paragraph) on a three-point scale, plus some limitation codes; a readability index; and a lexical density index. Also, an assessment of the consistency in language and notation across the articles is made.

4.3 Task specification: Step 2

To specify the task, the articles to evaluate must be determined. Many Wikipedia entries could be examined: as of 19 September 2017, 3679 statistical articles were within the WikiProject Statistics (WikiProject Statistics 2017). PDF versions of the actual articles evaluated can be obtained by contacting the authors.

We evaluate terms that are often an integral part of professional communication, such as consumer finance and workplace documents. Among these, we have chosen fundamental and important terms: two from descriptive statistics (“arithmetic mean”, “standard deviation”); one from graphical statistics (“histogram”); and two from inferential statistics (“confidence interval”; “standard error”) that are commonly misunderstood (Aliaga et al. 2010; Chance et al. 2004; Dunn et al. 2015; Fidler and Cumming 2005; Kaplan et al. 2009; Richardson et al. 2013; Vaux 2012).

We first evaluated the articles in August 2016. The evaluations were revisited in December 2016, and four of the five articles had been edited in that time. The article "Standard error" had received substantial improvements; "Arithmetic mean" had no changes; and minor changes had been made to the other articles. Interestingly, changes were mainly additions and minor changes, and very few cases of deletions were apparent.

Table 1 shows how these articles are rated in Wikipedia's rating system of quality and importance in the WikiProject Statistics, where the terms appear in an Internet search using DuckDuckGo (duckduckgo.com), plus other objective details from each article. DuckDuckGo was chosen as the search engine (rather than, say, Google), as searches are not personalized and hence the same results will be returned for all users.

Table 1 shows that the articles investigated have a substantial proportion of their references from refereed material, although we make no claim as to their relevance or usefulness. Perhaps more surprising is that the number of references for articles linked to the "mean" is so few. The opportunity to direct the enquirer to other material of value through "Further Reading" seems underutilized, and may represent a simple modification that adds value for time spent upgrading.

4.4 Methods: Step 3

Step 3 of Print's (1993) process requires a discussion of the methods used for gathering the identified information. Initially, two articles were examined independently by the authors, who then met to discuss how well the criteria captured the necessary features of the articles. This led to revisions of the criteria and rating system. Once the final rating system was developed, each team member made a final allocation of a rating (and, if necessary, limitation code(s)) to the two initial articles. These allocations were then discussed until consensus was reached among the team.

The team members then independently assessed the remaining articles, and met to discuss any discrepancies in the allocated ratings and explanations. Each team member justified

their allocated rating, and the reasons for the limitations. Discussions were held and coffee consumed until consensus was reached.

5. RESULTS AND ANALYSIS

Steps 4 and 5 of Print's model are data collection and observation, and analysis of the data respectively. In Section 5.1, we have gathered information suitable for comparison between articles, while Sections 5.2 and 5.3 discuss emergent issues and concerns.

5.1 Summative Outcomes

Because of the dynamic nature of Wikipedia articles, the ratings were allocated over a short time period, from 1 August 2016 to 5 August 2016. The results are shown in Table 2.

Clearly, all ratings are subjective, and readers may or may not agree with us. Nonetheless, the ratings represent a means for evaluating the Wikipedia articles on different dimensions at one point in time.

Whole articles were never rated more highly than the Preamble, and the Preamble never rated more highly that the first paragraph. Similarly, articles are less lexically dense than the Preamble, and the Preamble less dense than the first paragraph. Combined, these results suggest that the initial paragraphs are rated more highly and are more information rich. Readability results are more inconsistent; initial paragraphs range from 7 (lower secondary school/middle school) for "mean" to 15 (upper university level) for "confidence interval". Lexical density values are similar to those reported elsewhere for Wikipedia articles (Elia 2009; Analyze My Writing 2016).

Despite attempts to standardize the appearance of Wikipedia articles through templates, a lack of precision, poor pacing for concept development and a sense of being a perpetual first draft remain evident in articles considered. This is because in part an article will often reference another through an in-text hyperlink.

5.2 Limitations evident in the articles

The most common limitation referred to is the **display of graphics** (D). For example, every article is given this limitation for the whole article. Many graphics lack direct relevance or are of poor resolution. Examples are common in the histogram article: many of the images are fuzzy, the small fonts are very hard to read, and horizontal axes are often poorly labelled, if at all. Poor quality images are disappointing for an article about a graphical display.

Some images are ambiguous. The histogram article includes an example on aces served in tennis games; the histogram includes a label on the horizontal axis for "10 aces" that is located on the border between two histogram bars so that the location of an observation of 10 aces is unclear. The "arithmetic mean" article contains an image (Figure 1) with no axis labels and no explicit explanation of symbols used. For example, the symbol on the graph (σ) in Figure 1, which is never explained, may refer to the standard deviation (for which the symbol is commonly used) but the caption implies only the skewness differs between the two examples. In any case, the use of the diagram in an article on "arithmetic means" is questionable when a simpler diagram with less cognitive overload could be used.

A further example appears in the "confidence interval" article (Figure 2). The image has no context; the legend is for black (A) and brown (B) colors, yet the graph includes brown and white bars, with red line segments; and the caption describes the *bars* as symmetric, but almost certainly the comment is meant to refer to the red *line segments* (the confidence intervals). (As an aside, Weissgerber et al., (2015) contain criticisms of these types of plots in general.)

Some images are unsuitable, often because of recycling from other articles (where they may be suitable). For example, Figure 3 appears in the articles "standard deviation" and "standard error" (where the use of the commonly-used symbol for "standard deviation" is confusing); Figure 1 also appears in (and is more suitable for) the "log-normal" article.

Another commonly-applied limitation was that of **accuracy** (A), which encompasses errors, ambiguities, omissions and inconsistencies.

An example of an error appears in the "confidence interval" article, where the article repeatedly refers to confidence intervals as applying only to "experiments"; two examples appear in the first paragraph of the Preamble.

An example of an inconsistency appears in the "histogram" article. The first paragraph of the Preamble states that a histogram is "an estimate of the probability distribution of a continuous variable (quantitative variable)" (see MacGillivray et al. 2014, pp. 75–76), yet a later example (on the number of aces served in tennis) is for discrete quantitative data. A further example of inconsistency, in the context of notation, appears in the "standard deviation" article. The section "Assumptions and usage" states that

> The notation for standard error can be any one of SE, SEM (for standard error of *measurement* or *mean*), or S_E.

However, this notation does not describe the notation sometimes used in the article for the standard error of the mean ($SE_{\bar{x}}$), and the commonly-used $s_{\bar{x}}$ and s.e.($\bar{x}$) are not listed at all.

Some information is inappropriate. For example, the entire "standard error" article (every section and subsection, apart from the single Preamble paragraph) concerns the standard error of the mean.

Some articles are ambiguous. For example, the second diagram in the "confidence interval" article contains the expression "250 g ± 2.5 g" which is confusing (Motulsky 2015) and not explained. This interval does not refer to a confidence interval even though the image appears in the "confidence interval" article; the "2.5 g" actually refers to the population standard deviation.

Conceptual explanations (C) are generally acceptable for our intended purpose, but are lacking for the "mean". **Procedural explanations** (P) are sometimes lacking. For example, about 20% of the "histogram" article's words are spent explaining different methods for finding number of bins when constructing a histogram (after explaining there

is no “best” way), but this procedure is never demonstrated or used in any example to construct a histogram.

5.3 Overall comments

One difficulty with the “Arithmetic mean” article is locating it. “Mean” can wear many qualifiers (arithmetic mean; sample mean; geometric mean; etc.) and users may only know one of these as the “mean”, and may not even be aware that other types of “means” exist (Dunn et al. 2016b). However, the Wikipedia articles for “mean” (Mean 2016) and “sample mean” (redirected to “Sample mean and covariance”; Sample mean and covariance 2016) are about means in a general sense, and not the “sample arithmetic mean” as probably intended. In addition, both these articles are too technical for the intended audience. “Arithmetic mean” seems to be the appropriate article.

Erratic development also presents as a difficulty. The histogram article described above, where 20% of the words concern finding bins, is one example. The “standard deviation” article also has erratic development and lacks the consistency needed for building understanding. Much of the article is too theoretical and nuanced for our audience.

Notation may also be inconsistent. For example, the symbol σ is commonly used to represent the population standard deviation, but is also used to represent skewness (Figure 1). Sigma (σ) appears to be used for the standard error in that article, but really represents a standard deviation, while in the Standard Deviation article SD represents the standard deviation in one figure in that article, while σ is used in Figure 3 and in the article itself. These across- and within-article inconsistencies could prove frustrating for self-learners.

6. DISCUSSION AND CONCLUSION

Step 6 of Print’s model is to discuss the results. In offering these discussion points, we re-state that we know the purpose of Wikipedia is not to act as a learning resource, but we evaluate the five Wikipedia articles through this lens since many users use Wikipedia for this purpose, if only as an entry point.

Firstly, Wikipedia articles evaluated in this paper do not come recommended for learning by general users, though some initial paragraphs ("standard deviation" and "standard error") are suitable. The most common limitation code applied was D, suggesting poor presentations and displays. Contributing to this was the inappropriate reuse of images from other articles. Noticeably, no articles took advantage of the capabilities afforded by the online medium, such as animations and videos.

Disturbingly, the limitation of A (accuracy) was awarded six times (for "confidence intervals" and "histograms" for all components). This alone suggests that Wikipedia is not a recommended resource for self-learning. Indeed, problems with accuracy have far wider implications that just our context.

Despite the dearth of 3-ratings, a rating of 2 is common: for entire articles (2 articles), preambles (4) and first paragraphs (3). These, then, are deemed "suitable for self-learning when supplemented with authoritative resources". In summary, articles are best used to give the reader a broad overview of the topic, but cannot be recommended for deeper understanding due to errors and/or inconsistencies.

In a more formal learning context, Parry (2008) argues that because we live in an information age, the genuine self-learning user will need to be able to engage with a wide variety of information (instructors, textbooks, television, the Internet, and so on) and be critical of it. Teachers and academics therefore need to support students to be able to do this so they can develop into independent lifelong learners (Upchurch 2011; Parry 2008).

For this environment, we offer one, as yet untested, idea to achieve this. In using Wikipedia, it is important that relevant stakeholders (students; teachers) understand how Wikipedia knowledge is constructed so they understand the potential problems with the articles (Soylu 2009). A suitable class activity, then, could be to present small groups of students with Wikipedia extracts (whole article, preambles, or first paragraphs, depending on time, students, the topic and the situation), and have students critique the extract. Students could, for example, be asked to identify two good and two bad things about the extract, and to share these with the class. This activity brings Wikipedia explicitly into

the conversation, engages students, and provides an opportunity for instructors to help students identify shortcomings with the articles. As such, it may even act as a way for instructors to remind their students of the utility of more formal resources (such as textbook, class notes, etc.).

One further issue to emerge from this study is that the inconsistent language used in statistics (Dunn et al. 2016b) means that self-learners may find themselves reading an article that is inadvertently irrelevant. In this article, for example, we have identified how the general user may struggle to locate the correct article for "mean" on Wikipedia (which is probably "arithmetic mean"); similar struggles may occur with "correlation" or "correlation coefficient".

The discipline of linguistics reacted to the fluid narrative of Wikipedia by establishing "*Wikipedia* Update Project" in 2007, "to update Wikipedia articles focused on linguistics and languages" (Linguistics n.d.). This project is no longer active, but the evidence presented here suggests that similar focused attention could bear Wiki-fruit: for example, the "histogram" article has existed since 2001 and had 726 editors, yet still is a poor-quality article with inaccuracies and poor images. However, we see numerous challenges in advocating for this approach: statistics is so much a part of professional communication, and taught and used at so many levels, that writing an article that meets the needs of every user would be impossible, even if restricted to just self-learners. Similarly, since statistics is used in many disciplines, many people may believe (sometimes correctly, sometimes incorrectly) they have the expertise to edit articles, and will do so according to their own preferences. However, language and notation are notoriously inconsistent in statistics (Dunn et al. 2016b). This means that no agreed-upon notation and language will serve all purposes. Even if such notation and language could be agreed upon, and even if the Wikipedia articles were otherwise suitable for recommending to the self-learning user, their usefulness may be limited because the notation and language may be inconsistent with that used in other learning resources. As always, a further caveat with using Wikipedia articles is that they have the potential to

change, so that an article recommended one week may no longer be suitable the following week.

For all these reasons, we believe that the use of Wikipedia as a formal self-learning resource for the general user without further qualification is not to be recommended. However, given that the self-learner will continue to use Wikipedia this way, it is the quality of this qualification where bridges can be built into concept development through the material to be found in Wikipedia.

Considering this discussion, the following proposals have merit for consideration:

1. Wikipedia be actively discouraged for self-learning (for example, by using a classroom activity), except to give a broad overview of the topic, and to point users to more formal resource.

2. In more formal learning environments, teachers need to be explicit about the way Wikipedia is to be used as a learning resource for course content (for example, formal statements about the use of Wikipedia could be included in syllabus documents and course outlines).

3. Teachers and academics need to move away from minimal contact with Wikipedia towards constructive engagement (for example, deconstructing articles and identifying problems as formal class activities).

4. A simple way for the statistics community to make substantial improvements is to increase the quality of images used in articles.

It is presumptuous to expect an academic review of this type to effect change in the Wikipedia protocols currently in place, but certainly it can inform the prevailing view of the distance between the intent of Wikipedia as imagined by its creators and the way it is being used. While it has obvious advantages regarding in-text hyperlinks to other Wikipedia material, further use could be made of the platform to include animations and interactive elements that improve engagement with the self-learning reader.

While the perception that Wikipedia articles lack accuracy remains, it is unlikely the material will enjoy broad acceptance. A more visible presence for Wikipedia's existing article ratings, and any indication of improvement showing article development, will assist users in deciding about aspects of the articles' content. Indeed, such a positive direction for article development may garner the support of those with considerable expertise to contribute.

REFERENCES

Adafre, S. F., and de Rijke, M. (2005), "Discovering Missing Links in Wikipedia," in *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 90–97. ACM.

Alexa Internet (2017) *Alexa* [online]. Available at http://www.alexa.com/topsites

Aliaga, M., Cobb, G., Cuff, C., Garfield, J., Gould, R., Lock, R., Moore, T., Rossman, A., Stephenson, B., Utts, J., Velleman, P., and Witmer, J. (2010), "Guidelines for Assessment and Instruction in Statistics Education: College Report," Technical report, American Statistical Association.

Almeida, R., Mozafari, B., and Cho, J. (2007), "On the Evolution of Wikipedia," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Analyze My Writing (2016), *A Tale of Two Densities: Wikipedia vs. The Short Story* [online]. Available at http://www.analyzemywriting.com/Ideas/lexical_density_comparison.html

Arithmetic mean (2016). In *Wikipedia* [online]. Available at

https://en.wikipedia.org/wiki/Arithmetic_mean

Begeny, J. C., and Greene, D. J. (2014), "Can Readability Formulas be used to Successfully Gauge Difficulty of Reading Materials?" *Psychology in the Schools*, 51, 198–215.

Black, E. (April 19, 2010), "Wikipedia—The Dumbing Down of World Knowledge," *History News Network.*

Brown, A. R. (2011), "Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage," *PS: Political Science & Politics*, 44, 339–343.

Category: FA-Class Statistics Articles (2017) [online]. Available at: https://en.wikipedia.org/wiki/Category:FA-Class_Statistics_articles

Chance, B., delMas, R., and Garfield, J. (2004), "Reasoning about Sampling Distributions," in *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, pp. 295–323.

Distaso, M. W. (2012), "Measuring Public Relations Wikipedia Engagement: How Bright is the Rule," *Public Relations Journal*, 6, 1–22.

Dunn, P. K., Carey. M. D., Farrar, M. B., Richardson, A. M. and McDonald, C. (2016 a), "Introductory Statistics Textbooks and the GAISE Recommendations," *The American Statistician.* In press.

Dunn, P. K., Carey, M. C., Richardson, A. M., and McDonald, C. (2016b), "Learning the Language of Statistics: Challenges and Teaching Approaches," *Statistical Education Research Journal* [online], 15. Available at http://iase-web.org/documents/SERJ/SERJ15(1)_Dunn.pdf.

Dunn, P. K., Marshman, M., McDougall, R., and Wiegand, A. (2015), "Teachers and Textbooks: On Statistical Definitions in Senior Secondary Mathematics," *Journal of Statistics Education* [online], 23(3). Available at http://www.amstat.org/publications/jse/v23 n3/dunn.pdf

Elia, A. (2009), "Quantitative Data and Graphics on Lexical Specificity and Index Readability: The Case of Wikipedia," *RAEL: revista electrónica de lingüística aplicada*, 8, 248–271.

Fidler, F., and Cumming, G. (2005), "Teaching Confidence Intervals: Problems and Potential Solutions," in *Proceedings of the International Statistical Institute 55th Session*.

Giles, J. (2005), "Internet Encyclopaedias go Head to Head," *Nature* [online], 438, 900–901. http://dx.doi.org/10.1038/438900 a

Hartley, J. (2016), "Is Time Up for the Flesch Measure of Reading Ease?" *Scientometrics*, 107, 1523–1526.

Head, A., and Eisenberg, M. (2010), "How Today's College Students use Wikipedia for Course-related Research," *First Monday,* 15. doi:10.5210/fm.v15i3.2830

Julien, H., and Barker, S. (2009), "How High-School Students Find and Evaluate Scientific Information: A Basis for Information Literacy Skills Development," *Library & Information Science Research,* 31, 12--17.

Kaplan, J. J., Fisher, D. G., and Rogness, N. T. (2009), "Lexical Ambiguity in Statistics: What do Students know about the Words Association, Average, Confidence, Random and Spread?" *Journal of Statistics Education* [online], 17. Available at http://www.amstat.org/publications/jse/v17 n3/kaplan.html

Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B., Mytkowicz, T. (2007), "Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie," *Alt.CHI*, 2007. San Jose, CA; 2007 April 28 – May 3; San Jose, CA.

Lanier, J. (2006), "Digital Maoism: The Hazards of the New Online Collectivism," *The Edge*, 183, 30.

Linguist (n.d.). *The WikiProject: Linguistics* [online]. Available at http://linguistlist.org/projects/wikiproject.cfm

Linnarud, M. (1976), "Lexical Density and Lexical Variation—An Analysis of the Lexical Texture of Swedish Students' Written Work," *Studia Anglica Posnaniensia*, 7, 45–52.

MacGillivray, H., Utts, J. M., and Heckard, Robert F. (2014). *Mind on Statistics* (2 nd ed.), South Melbourne, Victoria: Cengage Learning.

Magnus, P. D. (2009), "On Trusting Wikipedia," *Episteme*, 6, 74–90.

McDougall, R. G., Flanders, M., Buchanan, R., and Lindsay, S. (2003), "Developing a Local Framework for Quality in an Online Learning Environment: A Case Study," in *Proceedings of the IFIP TC3/WG3.6 (International Federation for Information Processing) Conference: Quality Education @ a Distance*, pp. 93–100.

Mean (2016). In *Wikipedia* [online]. Available at https://en.wikipedia.org/wiki/Mean

Meriam Library, California State University-Chico. (2010), "Evaluating Information—Applying the CRAAP Test," *Evaluating information—Applying the CRAAP test* [online]. Available at http://www.csuchico.edu/lins/handouts/eval_websites.pdf

Motulsky, H. J. (2015), "Common Misconceptions about Data Analysis and Statistics," *British Journal of Pharmacology*, 172, 2126–2132.

Parry, D. (2008), "Wikipedia and the New Curriculum," *Science Progress* [online]. Available at http://www.scienceprogress.org/2008/02/wikipedia-and-the-new-curriculum/

Petrilli, M. J. (2008), "Wikipedia or Wickedpedia?" *Education Next* [online], 8, 87. Available at http://educationnext.org/wikipedia-or-wickedpedia/

President and Fellows of Harvard College (2016). *What's Wrong with Wikipedia?* [online]. Available at http://isites.harvard.edu/icb/icb.do?keyword = k70847&pageid = icb.page346376

Print, M. (1993), *Curriculum development and design* (2 nd ed.), Crows Nest, Australia: Allen & Unwin.

Read, B. (2006), "Can Wikipedia ever Make the Grade," *Chronicle of Higher Education*, 53, A31.

Richardson, A. M., Dunn, P. K., and Hutchins, R. (2013), "Identification and Definition of Lexically Ambiguous Words in Statistics by Tutors and Students," *International Journal of Mathematical Education in Science and Technology*, 44, 1007–1019.

Sample mean and covariance (2016) In *Wikipedia* [online]. Available at https://en.wikipedia.org/wiki/Sample_mean_and_covariance

Schönhofen, P. (2006), "Identifying Document Topics using the Wikipedia Category network," *Web Intelligence and Agent Systems: An International Journal*, 7, 195–207.

Search Engine People (2015) *10 Most Notorious Wikipedia Editing Scandals* [online].

Available at http://www.searchenginepeople.com/blog/most-notorious-wikipedia-scandals.html

Selwyn, N., and Gorard, S. (2016), "Students' use of Wikipedia as an Academic Resource—Patterns of Use and Perceptions of Usefulness," *The Internet and Higher Education*, 28, 28–34.

Soylu, F. (2009), "Academics' Views on and uses of Wikipedia," *Journal of Communication, Culture and Technology*, 9. Available at http://www.gnovisjournal.org/2009/05/13/academics-views-and-uses-wikipedia/

Strube, M., and Ponzetto, S. P. (2006), "WikiRelate! Computing Semantic Relatedness using Wikipedia," *AAAI*, 6, 1419–1424.

Thomas Jefferson University. (2010), "Cancer Information on Wikipedia is Accurate, but not very Readable, study finds." *Science Daily* [online]. Available at www.sciencedaily.com/releases/2010/06/100601114641.htmhttp://www.sciencedaily.com/releases/2010/06/100601114641.htm

http://www.sciencedaily.com/releases/2010/06/100601114641.htm

To, V., Fan, S., and Thomas, D. (2013), "Lexical Density and Readability: A Case Study of English Textbooks," *Internet Journal of Language, Culture and Society*, 37, 61–71.

Tynan, D. (2008), "The Truth is out there… Somewhere," *US Airways Magazine*, p. 42.

Upchurch, J. (2011), *Examining Wikipedia's Value as an Information Source Using the California State University-Chico Website Evaluation Guidelines* [online]. Available at http://files.eric.ed.gov/fulltext/ED522722.pdf

Vaux, D. L. (2012), "Know When your Numbers are Significant," *Nature*, 492, 180–181.

Viégas, F. B., Wattenberg, M., and Dave, K. (2004), "Studying Cooperation and Conflict between Authors with History Flow Visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 575–582.

Waller, V. (2011). "The search queries that took Australian Internet users to Wikipedia" *Information Research*, **16**(2) paper 476. [Available at http://InformationR.net/ir/16-2/paper476.html]

Waters, Neil L. (2007), "Why you can't Cite Wikipedia in my Class," *Communications of the ACM*, 50, 15–17.

Weissgerber, T. L., Milic, N. M., Winham, S. J. and Garovic, V. D. (2015), "Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm," *PLOS Biology*.

https://en.wikibooks.org/wiki/Statistics

Wikipedia. (2016). In *Wikipedia* [online]. Available at http://en.wikipedia.org/wiki/Wikipedia

Wikipedia: Featured Articles (2017) [online]. Available at https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

Wikipedia: Featured Article Criterion (2017) [online]. Available at https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Wikipedia: Version 1.0 Editorial Team. (2016) [online]. Available at https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team

WikiProject Mathematics (2016) [online]. Available at https://en.m.wikipedia.org/wiki/Wikipedia:WikiProject_Mathematics/Wikipedia_1.0/Assessment

WikiProject Statistics. (2017) [online]. Available at https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Statistics

Yasseri T., Sumi R., Rung A., Kornai A., and Kertész J. (2012), "Dynamics of Conflicts in Wikipedia," *PLoS ONE*, 7: e38869. doi: 10.1371/journal.pone.0038869

Zeng, H., Alhossaini, M. A., Ding, L., Fikes, R., and McGuinness, D. L. (2006), "Computing Trust from Revision History," in *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services.* Available at http://ebiquity.umbc.edu/_file_directory_/papers/302.pdf

Table 1. Information about the Wikipedia articles, as of 05 August 2016.

Article	Wikipedia quality	Wikipedia importance	DDG search rank	Average no. daily page views	No. editors	Year of first edit	Refereed refs	Total no. references	External links	Further reading
Mean	B	Top	2 nd	1074	1044	2001	4	5	4	0
Sample mean	Start	Mid	1 st	137	69	2007	2	3	0	0
Arithmetic mean	C	Top	1 st	846	421	2001	3	4	3	1
Standard deviation	C	Top	1 st	8209	2024	2001	10	15	4	0
Standard error	C	Top	1 st	2002	302	2004	10	10	0	0
Confidence interval	C	High	2 nd	2692	607	2003	25	30	7	12
Histogram	B	Top	1 st	1506	726	2001	16	19	10	1
Medians	**C**	**Top**	**1 st**	**1506**	**607**	**2003**	**10**	**10**	**4**	**0**

Notes: Most information as reported by Wikipedia Project Statistics (2016) or found in the articles. DuckDuckGo (DDG) search results found with duckduckgo.com with compound terms placed in quotation marks (for example, "arithmetic mean"). Information for "mean" and "sample mean" (which redirects to "Sample mean and covariance") included because these may be the search terms that self-learners will use (see Section 5.2). All average daily page views are for the 90 days up to 5 August 2016. Article templates usually have the sub-heading 'Further Reading', but the article for Confidence Interval has the sub-heading 'Bibliography' instead.

Table 2: The information for each article: for the first paragraph, the preamble and the whole article.

	First Preamble paragraph						Preamble							Article					
Article title	**Rating**	**Readability**	**Lex. den. (%)**	**No. words**	**No. sentences**	**No. images**	**Rating**	**Readability**	**Lex. den. (%)**	**No. words**	**No. sentences**	**No. paragraphs**	**No. images**	**Rating**	**Readability**	**Lex. den. (%)**	**No. words**	**No. sentences**	**No. images**
Arithmetic mean	2 C	6.8	70	40	11	0	2 C	10.7	68	125	17	4	0	2CD	10.3	53	1811	366	2
Standard deviation	3	12.4	74	42	4	2	2D	16.5	57	267	16	4	2	1PD	8.9	34	4601	822	7
Standard error	3	12.7	78	23	2	1	2D	12.0	54	102	9	3	1	1D	9.9	48	2064	411	3
Confidence Interval	2 A	15.2	70	76	7	1	2AD	14.5	58	231	25	5	0	2AD	9.9	35	4934	1184	3
Histogram	2 A	12.6	81	53	6	1	1AD	10.5	60	248	26	8	1	1APD	7.2	49	1936	533	15
Column mean:	2.4	11.9	75	47	6	1	2.2	12.8	59	195	19	5	1	1.4	9.2	44	3069	663	6
Column median:	2	12.5	74	42	6	1	2	12.0	58	231	17	4	1	1	9.6	46	2064	533	3

Notes: Ratings are based on a three-point ordinal scale: 1 (Don't recommend: not suitable for a self-learners); 2 (Suitable for self-learners when supplemented with authoritative resources); 3 (Recommended for self-learning without further qualification). The limitation codes are explained in Section 4.2. Readability is the Gunning Fog index. Readability, lexical density (Lex. den.), word and paragraph counts are measured using textalyser.net. For the whole article, the URL was provided to textalyser.net, and so images and navigational elements are included. For the Preamble, the text (including images but no navigation elements) were cut-and-paste into the textalyser.net webpage.

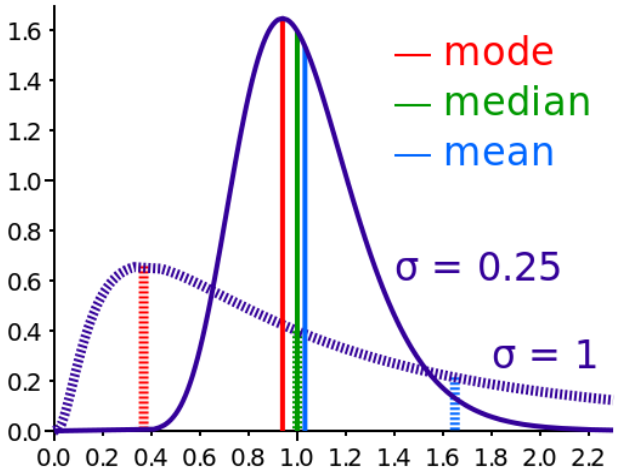

Figure 1. The diagram appearing in the Wikipedia article for "Arithmetic mean". The original caption is "Comparison of mean, median and mode of two log-normal distributions with different skewness."

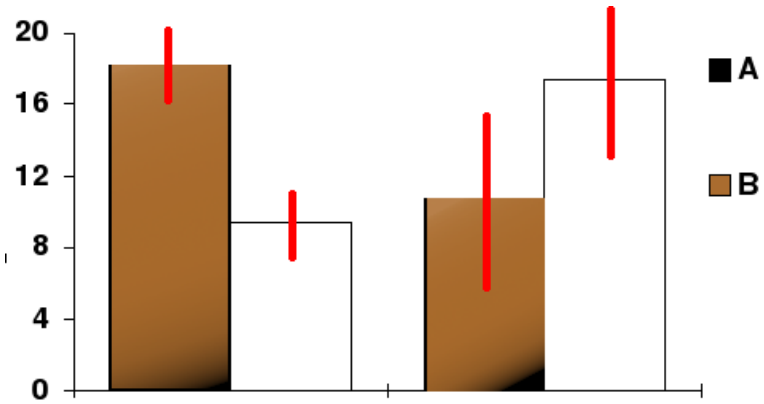

Figure 2. The first diagram in the Wikipedia article "confidence interval". The original caption is: "In this bar chart, the top ends of the bars indicate observation means and the red line segments represent the confidence intervals surrounding them. Although the bars are shown as symmetric in this chart, they do not have to be symmetric."

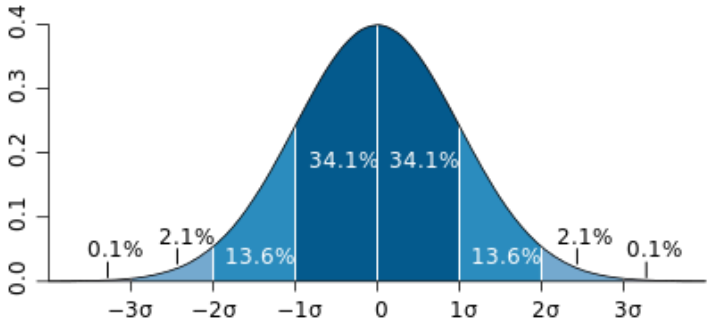

Figure 3. The diagram in the Preamble for the Wikipedia articles "standard deviation" and "standard error". In the "standard deviation" article, the original caption is "A plot of a normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation – See also: 68–95–99.7 rule". The original caption in the "standard error" article is: "For a value that is sampled with an unbiased normally distributed error, the above depicts the proportion of samples that would fall between 0, 1, 2, and 3 standard deviations above and below the actual value."