Early Statistical Findings and Authorship Misattribution: An Unsystematic Review of the Literature

Nigel Smeeton

Centre for Research in Public Health and Community Care, University of Hertfordshire, United Kingdom (n.smeeton@herts.ac.uk)

For historical accuracy and in fairness to the authors of original work, it is right and proper that acknowledgment should be given for their contribution to scientific advances. I first encountered authorship misattribution in the statistical literature in the early 1980s. In 1971, Richard Light had reported an apparently new measure of observer agreement known as the conditional kappa coefficient in *Psychological Bulletin*. Shortly afterwards, I came across a copy of *Fingerprints* (Galton, 1892), in which Francis Galton described a ‘centesimal scale’ which he used to assess the similarity of the patterns on corresponding fingers of the left and right hands. Comparison with Light’s paper shows that Galton had proposed the same measure, which I pointed out in a letter to *Biometrics* in 1985.

Since then, I have seen other examples of author misattribution in a statistical context. These have been reported as individual cases. However, I have yet to discover a collection of reports of misattribution obtained through a literature review.

Methods

Reports of misattribution were identified from key papers relevant to my academic interests, published letters regarding misattribution, and citations of such letters.

For clarification, the first known document containing an innovative finding is referred to here as the 'original publication', and the publication currently regarded as the source material as the 'conventional source'.

This work reports on misattribution in early statistical work; as discussed later this links in with my interest in Victorian fiction. The search was therefore restricted to original publications that appeared no later than 1914. It was assumed that the authors who incorrectly regarded their work as original research published in good faith and were unaware of previous publication. To make this assumption realistic, reports of misattribution were restricted to those for which the time between the original publication and the conventional source was at least 25 years. Authors of conventional sources were given credit for any evidence of a search of the literature available to them.

Results

In addition to my own discovery, three reports of misattribution were identified. Two were found in papers relevant to my academic interests, and one from the citations of a published letter regarding authorship misattribution. Further details are as follows:

Kappa Statistic for Observer Agreement

In its simplest form, two judges consider *n* items and allocate each into one of two categories. The number of items for which there is agreement on Category 1 is written as n_{11}, and the number for which there is agreement on Category 2 is n_{22}. Where the assessors disagree, the corresponding numbers are n_{12} (Judge 1 allocates to Category 1, Judge 2 to Category 2) and n_{21} (Judge 1 allocates to Category 2, Judge 2 to Category 1). The numbers for each combination of categories are shown in Table 1, with row totals indicated by $n_{1.}$ and $n_{2.}$ and the column totals by $n_{.1}$ and n$_{.2}$.

	Judge 2 – Cat 1	Judge 2 – Cat 2	Total
Judge 1 – Cat 1	n_{11}	n_{12}	$n_{1.}$
Judge 1 – Cat 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Table 1– Allocation of items to one of two categories by two judges.

What is the level of agreement between the judges? A straightforward answer is the proportion of items on which the two judges agree, i.e. $p_o = (n_{11}+ n_{22})/ n.$ However, some of this agreement will have occurred on the grounds of chance. For instance, if a fair coin is repeatedly tossed to make allocations based on the outcomes Head and Tail the expected proportion of agreement is 0.5. Assessment of the true level of agreement between the judges requires a measure that is chance-corrected.

The kappa statistic κ is based on the proportion of items assigned to the same category under complete agreement (i.e. 1), the expected proportion of agreement, p_e, and the observed proportion of agreement, p_o. It is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where $p_e = \frac{(n_{1.} \times n_{.1})+(n_{2.} \times n_{.2})}{n^2}$.

It is widely believed that the psychologist Jacob Cohen first proposed the kappa statistic in the journal *Educational and Psychological Measurement* in 1960. His paper starts with a review of the earlier literature that covers the previous 20 years. Unsurprisingly, Cohen failed to discover that this measure of agreement had been presented by Myrick Doolittle as an ‘association ratio’ at a meeting of the Philosophical Society of Washington held in 1887. Timothy Armistead pointed out this omission in *The American Statistician* in 2016. In addition, Armistead discussed links between association measures and Bayes’ Theorem.

Conditional Observer Agreement

An alternative form of the kappa statistic is based on the maximum possible proportion of agreement (p_{max}) obtainable given the row and column totals. This conditional kappa statistic is calculated as $\frac{p_o - p_e}{p_{max} - p_e}$.

As mentioned in the introduction, this statistic is generally attributed to Light. Although Light interspersed his paper with references to previous work, he did not identify Galton’s centesimal scale of 1892, which amounts to the same measure.

Francis Galton was interested in the degree of matching between pattern types on the fingers of the left and corresponding right hands. He had classified the papillary ridges found on the fingers into arches, loops and whorls. Galton’s observations for the ring-fingers of 100 individuals according to the presence or absence of a whorl are summarized in Table 2.

	Right – Whorl	Right – Other	Total
Left – Whorl	26	5	31
Left – Other	19	50	69
Total	45	55	100

Table 2–Presence or absence of a whorl on left and corresponding right ring-fingers.

Galton reasoned that given the row totals, the number in the whorl/whorl cell cannot exceed 31. The observed number is seen to be 26 and he calculated the number expected by chance as (45 × 31)/100 or 14 to his approximation. He proposed a centesimal scale for chance corrected agreement having 0° as “no relationship” and 100° as “utmost feasible likeness”. For the table above, Galton calculated the centesimal scale value as:

$$\frac{26-14}{31-14} \times 100^{\circ} = 71^{\circ}.$$

Use of Light’s conditional kappa formula involves the same calculation and gives the equivalent answer.

Interestingly, Galton expressed “grave objection” regarding use of his centesimal scale, stating that it requires the assumption that the average proportions for the arch, loop and whorl categories are preestablished in the population, e.g. that in general 45 percent of right ring-fingers have a whorl. This perceived limitation has tones of the controversy regarding the use of fixed marginal totals in Fisher’s exact test, illustrated by George Barnard’s remarks in a 1945 issue of *Nature*.

Wilcoxon-Mann-Whitney Test

Suppose that two samples have been collected and there is interest in whether the two population distributions are the same or differ in respect of their medians. If the distributions are skewed, a test based on the ranks of the data is appropriate. Frank Wilcoxon is generally given the credit for the test explained below, which appeared in *Biometrics* in 1945. This procedure was shortly followed by an equivalent test developed independently by Henry Mann and D. Ransom Whitney, published in a 1947 issue of *The Annals of Mathematical Statistics*. It is therefore conventional to refer to the tests jointly as the Wilcoxon–Mann–Whitney (or WMW) test.

In the Wilcoxon formulation of the WMW test, the two samples are combined, and the data put in order of size and ranked. The original samples are then separated, with each rank being attached to the corresponding observation. For identical populations, each sample should contain a similar mixture of observations of high, low, and intermediate rank. If the population medians differ, low ranks will predominate in one sample and high ranks in the other. The sum of the ranks of the observations in the smaller group are used to assess the significance of the difference between the medians of the two populations. Relatively low or high values for this sum provide evidence against the two populations being the same.

Neither of the WMW test papers include a review of the relevant literature. It was subsequently found that, writing in the German language, Gustav Deuchler had described a corresponding procedure in the journal *Zeitschrift für Pädagogische Psychologie und Jugendkunde* in 1914. William Kruskal reported Duechler’s earlier work in the *Journal of the American Statistical Association* in 1957, helpfully providing an English translation of the relevant section of the source article, along with an explanation of its equivalence.

One-sample Runs Probabilities

Suppose that a coin is tossed and the outcomes (Head – H or Tail – T) are recorded as a sequence. A run is regarded as a repetition of outcomes of the same type so, for example, HHTTTTHH consists of three runs. What are the probabilities of obtaining a particular number of runs?

Denote the number of tosses in the sequence of observations by *N*, with *m* heads and *n* tails. Let *r* indicate the number of runs. The cases *r* odd and *r* even are considered separately. For *r* odd, if *s* is such that *r* = 2*s* + 1, the probabilities can be written as:

$$\frac{\binom{m-1}{s-1}\binom{n-1}{s}+\binom{m-1}{s}\binom{n-1}{s-1}}{\binom{N}{m}}$$

If *r* is even, with *r* = 2*s*, the probabilities are:

$$2\times\frac{\binom{m-1}{s-1}\binom{n-1}{s-1}}{\binom{N}{m}}$$

These formulas are widely believed to originate from work published in *The Annals of Eugenics* by WL Stevens in 1939. The list of references in this article is very short and does not mention earlier publications. The truth is that the Rev William Whitworth stated these results as unexplained answers to Exercises 193 and 194 in the 4th edition of his text *Choice and Chance* (Whitworth, 1886). Solutions followed in a separate Deighton Bell publication that appeared in 1897.

WOULD IT BE POSSIBLE FOR WHITWORTH'S SOLUTIONS TO BE INSERTED ABOUT HERE, PLEASE?

David Barton and Florence David, in their *Biometrika* paper on multiple runs, published in 1957, cited Whitworth's text as the oldest known source of the two-category formulas. However, they were doubtless acquainted with Whitworth's solutions too, as they described their approach as "following Whitworth".

Discussion

To my knowledge, this paper contains the first collection of reports of authorship misattribution. From my experience of systematic reviewing, this study has important limitations. The search 'strategy' was highly unsystematic and largely limited to my own statistical research and literary interests. Post-1914 original sources, and reports where the conventional source followed the original publication within 25 years were not considered. Given that several instances were uncovered using this far from perfect process, a broader investigation of authorship misattribution is needed.

The strength of this review is that the reports found provide some pointers as to how to reduce the chance of overlooking important previous work.

Look Beyond the Peer Reviewed Literature

Two of the four works (Galton, Whitworth) were published as books with limited circulation; these are unlikely to have been peer reviewed. Editors have for many years rightly aimed to publish work of the highest quality. One of the tools used in their search is professional peer review. Superficially promising manuscripts may contain little of substance or, to borrow from Shakespeare’s play *The Merchant of Venice*: “all that glisters is not gold” (Act II, Scene VII).

However, the aim of a literature review is to cast the net as far as possible and identity relevant material whether good or not so good. Regarding the so called gray literature (internal reports, etc.), to take from *The Merchant of Venice* again: “but thou, thou meagre lead, which rather threatens than doth promise aught” (Act III, Scene II) should be on the radar.

Defunct Journals

Journals that are no longer in existence, particularly those that disappeared many years ago, are likely to be missed by bibliographic databases. The journal *Bulletin of the Philosophical Society of Washington* ceased continuous production in 1910 (Doolittle) and *Zeitschrift für Pädagogische Psychologie und Jugendkunde* has not been published since 1944 (Deuchler).

Language Counts

Deuchler published in the German language rather than in English. Researchers in the United States and Britain, for example, may not have been familiar with the literature produced in other languages. Today, the bibliographic database SciELO eases this task. It includes many periodicals that publish in non-English languages, a particular strength being the Spanish and Portuguese language journals based in Latin America.

Expect the Unexpected

Whitworth started out as a schoolteacher but for almost 40 years, he was a priest in the Church of England. That he produced scientific publications may today seem rather surprising. From a reading of Barton and David's *Biometrika* paper, they believed that the formulas presented by Whitworth were unlikely to have been new to him. There was no evidence for an earlier source, so they possibly thought that given Whitworth's position in the Church this statement would go unquestioned.

What Barton and David appear to have overlooked, however, is that Whitworth's academic background was in mathematics and he spent much of his 'free' time pursuing his interests in this field. He was a founding editor of the now defunct journal *Messenger of Mathematics* and was the author of numerous papers, along with several texts. Karl Pearson contributed to the *Messenger of Mathematics* so he at least must have known of Whitworth.

A detailed account of Whitworth's life was presented by Joseph Irwin at a 1967 meeting of the Royal Statistical Society (RSS), which was published in the Society's *Journal* later that year. The discussion that followed Irwin's account indicates that although acknowledging Whitworth's competence as a mathematician some had doubts as to whether he was the originator of the runs probabilities formulas.

Somewhat surprisingly, the RSS *Journal* report of the 1967 meeting indicates that Florence David had switched her view from that expressed in 1957. She spoke up strongly for Whitworth stating: "I myself would give him credit with runs distributions – Problems 193, 194 and 687. He was the first to give a general solution in a manipulable combinatorial form” and “no one appears before Whitworth to have written out explicitly the distribution of the number of runs. He was the first, I think”.

The moral of the story is that in searching for the development of statistical ideas prior to 1900, there is good reason to look beyond conventional sources such as statistical publications. As pointed out by the historian Theodore Porter, in the nineteenth century individuals working in different fields crossed paths more than is the case today. The statistician Karl Pearson was acquainted with the novelist and poet Thomas Hardy and the playwright George Bernard Shaw. This facilitated the cross-fertilisation of ideas. In Hardy’s novel *A Laodicean*, published in 1881, a character discusses probability and runs of events of the same type in the playing of roulette, stating to a financially reckless young man: “these runs of luck will be your ruin”. Randomness also features in some of Hardy’s poetic work. For instance, his poem *Hap*, written in 1866, contains the lines: “Crass Casualty obstructs the sun and rain, and dicing Time for gladness casts a moan”.

Concluding Comment

This study has only scratched the surface of a possibly significant phenomenon. Further clear examples of misattribution need to be reported and cases for which misattribution is a possibility should be examined. An accessible misattribution database is required in order to reduce the risk of misattribution in the future.

Further Reading

Doolittle, M.H. 1888. Association ratios. *Bulletin of the Philosophical Society of Washington*, 10: 83-87 & 94-96. *https://www.biodiversitylibrary.org/item/245135#page/181/mode/1up*

Galton, F. 1892. *Fingerprints*. London: Macmillan. *http://galton.org/books/finger-prints/galton-1892-finger-prints-1up.pdf*

Kruskal, W.H. 1957. Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, 52(3):356-360.

Porter, T.M. 2004. *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton: Princeton University Press.

Whitworth, W.A. 1886. *Choice and Chance. 4th edition*. Cambridge: Deighton Bell and Co. *https://downloads.tuxfamily.org/openmathdep/algebra/Choice_and_Chance-Whitworth.pdf*