NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

01-24

Looking Into the i of the Storm: A Study of the Mid-1880's Indices of Finley, Gilbert, Peirce and Doolittle and their Place in Contingency Table Analysis

Eric J. Beh

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia T: +61 2 42215076. E: karink@uow.edu.au

Looking Into the *i* of the Storm:

A Study of the Mid-1880's Indices of Finley, Gilbert, Peirce and Doolittle and their Place in Contingency Table Analysis

Eric J. Beh

National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia
&
Centre for Multi-Dimensional Data Visualisation (MuViSU), Stellenbosch University, South Africa
Email: ericb@uow.edu.au

Abstract

One of the first serious attempts to quantify how two categorical variables of a contingency table are associated was undertaken in the meteorological literature during the mid-1880's; 15-20 years prior to the significant contributions of Sir Francis Galton and Karl Pearson. Such work was motivated by the four-tornado data sets that Seargent John Park Finley collected and published in 1884; these data sets were analysed in the same fashion we analyse 2×2 contingency tables today. These data sets proved to be controversial, but it was the index he developed which drew most of the attention, an index designed to help verify the accuracy of the tornado predictions he made. Immediately following Finley's work, emendations of his index were proposed all of which involve ideas and concepts that pre-date the development of the analysis of contingency tables whose origins are universally credited to Galton, Pearson and others. Therefore, this paper provides an overview of Finley's data and the indices developed to give historical context to their place in the origin story of contingency table analysis.

Keywords

Association, 2×2 contingency tables, Finley's data, contingency, tornado, chi-squared statistic

1 Introduction

1.1 Practitioners of the Contingency Table: Pre-1900

The contingency table remains one of the most ubiquitous ways of displaying data, especially data of a categorical nature; these categories may exist as "natural" distinctions of traits or arise by converting numerical data to interval data. Stigler (2002) provides an excellent historical account of the contingency table starting with Sir Francis Galton (1822 – 1911) but does point out that the 2×2 contingency table has links to logic and probability dating back to the time of Aristotle. He then notes the early contributions of Karl Pearson (1857 – 1936), George Udny Yule (1871 – 1951) and Maurice Bartlett (1910 – 2002). Agresti (2013, Chapter 17) also provides a historical tour of categorical data analysis with its origins being Pearson, Yule and then Sir Ronald A. Fisher (1890 – 1962). While such contributions date back to the early 20th century, the contingency table has been used extensively for (at least) the past 400 years in a variety of economic, sociological and health contexts. Some of the more well-known consumers of the contingency table during this time include the following:

- The father of demography, John Graunt (1620 – 1674) published a book in 1662 titled *Natural and Political Observations, Mentioned in a following Index, and made upon the Bills of Mortality*. On pages 71 to 76 (inclusive) of this book, Graunt presents a series of contingency tables formed from cross-classifying, often by gender, the number of christened, married or buried (not deaths) recorded across the parishes of London with the year in which these occurred (ranging from 1569 to 1661, depending on the data)
- In 1797, Sir Frederick Eden (1766 – 1809) wrote a three-volume set of books spanning more than 1700 pages titled *The State of the Poor* that presented a wide range of details and counts across the England. Volume 2 gave parish-specific information on a range of social and financial issues that were tabulated in the form of a contingency table. For example, Eden (1797, Vol 2, p. 25) gives a table that cross-classifies the number of baptisms, burials and marriages (as well as records of household finances) by year during the period 1680 – 1795. Eden (1802, p. 45) also tabulated the local and foreign tonnage arriving and leaving the ports of England and Wales between 1770 and 1782.
- Belgium statistician and sociologist, Adolphe Quetelet (1796 – 1874) used contingency tables for better understanding the "social physics" of Holland (a term that would be replaced with "sociology"). For example, Quetelet (1835, p. 8) presented in French a contingency table that summarised the number of deaths in France between 1826 and 1831 (inclusive) due to a range of causes (including murder, guns, pistols, swords, knives, strangulations, drownings, kicks and blows, fire and unknown); this data would also appear on page 6 of his 1842 book (in English). An interesting discussion of Quetelet's place in statistics can be found in Hankins (1908, Chapter 2) with his remaining chapters discussing other interesting aspects of Quetelet.
- Someone for whom Quetelet was very influential was Florence Nightingale (1820 – 1910), the "Lady with the Lamp" of England. Jahoda (2015) discusses that Nightingale certainly knew of Quetelet and his work. The pair met each other in person only once, at the 1860 International Statistical Congress (Diamond & Stone, 1981, p. 66) that was held at King's College, London on the 16th of July. Nightingale's education was "above the normal for a

young woman in the nineteenth century England" having been surrounded by "many of the young lions of British science" (Diamond & Stone, 1981, p. 67) with Nightingale drawn towards mathematics. As such, Nightingale's analytical and numeracy skills aided her greatly and helped to forge her path and she used contingency tables extensively. For example, Nightingale (1858a, p. 15) cross-classified the years of experience of soldiers who took part in the Crimean War by the units they belonged too, including the number of soldiers that were made invalid in each unit. She also cross-classified the age of the soldiers in the British Army with how many in each division had died at home (Nightingale, 1858b, p. 9).

This, of course, is only a very short list of examples of how contingency tables have been used during the 17th, 18th and 19th centuries. More detail on these and other examples of early contingency tables can be found in, for example, Stigler (2002) and Beh & Lombardo (2014, pp. 54 – 61).

Briefly, the origin story of contingency table analysis often begins with Galton (1892). He analysed the fingerprint data from 105 pairs of twin brothers and is important not just because of the impact it had on contingency table analysis (which we talk more on in Section 4.4) but that it also was aided by his earlier description of correlation (Galton, 1888). Although it is the development of Pearson's (1904) chi-squared statistic that is the methodological beginning of contingency table analysis. In his 1904 paper, Pearson discusses the idea of two random variables from an underlying normal distribution being categorised and cross-classified to form what we now know to be a two-way contingency table. Pearson demonstrated with immense technical skill and rigour the derivation of what we know to be his chi-squared statistic and its links with Galton's (1888) description of correlation; in the broadest terms, Pearson (1900) derived the chi-squared statistic for univariate data while Pearson (1904) derived the statistic for bivariate data in the form of a two-way contingency table. Further discussions on the pre-history of Pearson's chi-squared statistic, including a detailed account of some of its predecessors, can be found in Lancaster (1966). One that should not be neglected from any discussion on the early period of contingency table analysis is George Udny Yule. He was very much interested in examining the technical and practical implications of correlation, association, and its relationship to contingency tables; see, for example Yule (1900, 1903) with much of his work during the mid-1890's furthering the discussion of correlation that Galton and Pearson had made. A very interesting account of the life and career of George Udny Yule and his links to Pearson and his work on correlation and association was given in by Kendall (1952).

It is the work of Galton, Pearson and Yule that would spark a whole new world of statistical theory and application including for the analysis of the contingency table. Although, one may find that the seeds of some of their ideas were planted half a world away and at least a decade earlier and it is these seeds that we shall focus our discussion on.

1.2 Predecessors of Galton and Pearson: Finley, Gilbert, Peirce and Doolittle

While the legacy left by Galton and his successors to contingency table analysis (and, more generally, to categorical data analysis) are deservingly still being felt amongst the statistical and her allied communities, there is an argument to be made for far more attention to be given to those whose efforts are still overlooked in favour of these early pioneers. We speak specifically here of the efforts of a group of four American meteorologists/scientists in the mid-1880's who developed a range of indices to help verify the accuracy of tornado predictions in central USA. The first of these indices was published by Seargent John Park Finley (1854 – 1943) in 1884 who, while working for the US Army Signal Service, gave a set of four data sets that is a cross-classification of the number of successful and unsuccessful predictions of tornadoes he made with the number tornadoes that occurred and did not occur. Accompanying his data, Finley (1884b) also proposed an index for verifying the accuracy of his predictions. We say at the outset that this index is fundamentally flawed (something we discuss at length below) however its importance stems from the attention that quickly followed; attention that focused on the development of improved indices and the initiation of statistical concepts that lie at the heart of contingency table analysis today. The contributors of these developments are very rarely, if ever, given full recognition outside of the meteorology communities. Even within these communities it is their meteorological contributions that is the focus of much of their attention and not their statistical contributions. Therefore, this paper will discuss Finley's tornado data, given by Figure 1, the index he developed for studying this data and the flurry of activity that quickly followed in 1884 and 1885. We shall be reviewing the indices proposed and discussed by Finley (1884b) and three of his successors – Gilbert (1884), Peirce (1884) and Doolittle (1885) – and the place they hold in origin story of contingency table analysis prior to the contributions of Galton, Pearson, Yule and others. This review will span the following eight sections. We start with a brief overview of the development of tornado observations (Section 2). Section 3 discusses the data Finley (1884b) collected and its presentation as a 2×2 contingency table formed from the cross-classification of the variables *Predicted* and Occurrence, these being the number of tornadoes Finley predicted (or not) and the number of tornadoes that occurred (or not). Section 3 will also describe the index Finley proposed for assessing the accuracy of his predictions. We shall start with the index of Gilbert (1884) in Section 4 and his proposal of e, a quantity that is equivalent to the expected frequency of the (1,1)th cell under complete independence (between *Predictions* and *Occurrence*). Following Gilbert's (1884) index is the index proposed by Peirce (1884) and the index of Doolittle (1885); these will be described in Section 5 and Section 6, respectively. Additional discussions on the relevance of this early work to the foundations of contingency table analysis established by Galton, Pearson and others will be made throughout Sections 3 to 6 (inclusive) as well as how these indices and their contributions fit within the analysis of contingency tables today. Section 7 will provide an in-depth assessment of the features of each index by determining their bounds, linkages, and behaviour for changes in the cell frequencies. Section 8 provides further practical matters outlined by Curtis (1887) concerning the analysis of Finley's tornado data and the development of the various indices. Some final remarks will be made in Section 9.

1.3 50 years of Contingency Table Analysis

While the focus of this paper is to look back at its pre-history in the mid-1880's it is also interesting to view the progress that has been made to contingency table analysis. So, this section provides a brief overview of the last 50 years of its development.

While the use of contingency tables in the description of social, economic and health conditions over the past few centuries is quite extensive, the development of tools and techniques for analysing them is quite poor prior to the 20th century. Before the 1900's the early attempts at quantification for contingency tables were largely confined to discussing raw counts, proportions or ratios. It was the work of Galton, and especially Pearson, Yule and other successors that would generate new ways of analysing contingency tables. The contributions made in the first half of the 20th century now means that the number of methodologies with a focus on contingency table analysis and, more generally, categorical data analysis, is, for lack of a better word, extensive. For example, at the risk of omitting entries, the following is a non-exhaustive list of 42 books published over the past 50 years that delve into the analysis of categorical data and/or contingency tables; Plackett (1974), Gokhale & Kullback (1978), Haberman (1978, 1979), Wrigley (1985), Wickens (1989), Andersen (1994, 1997), Blasius & Greenacre (1998), Everitt (1992), van de Geer (1993a, 1993b), Clogg & Shihadeh (1994), Le (1998, 2009), Lloyd (1999), Friendly (2000), Leonard (2000), Rayner & Best (2001), Simonoff (2003), Stokes, Davis & Koch (2003), van der Ark, Croon & Sijtsma (2005), Congdon (2005), Agresti (2007, 2010, 2013), Fienberg (2007), Powers & Xie (2008), Tutz (2011), Tang, He & Tu (2012), Kateri (2014), Sutradhar (2014), Bilder & Loughlin (2015), Friendly & Meyer (2018), Fagerland, Lydersen & Laake (2017), Upton (2017), Rudas (1998, 2018) and Azen & Walker (2021). One can also include in this list the collection of published works by L. A. Goodman; see Goodman & Kruskal (1979) and Goodman (1978, 1984)[1]. For some excellent discussions on the history of the contingency table and its pioneers the interested reader is directed to Killion and Zahn (1976), Stigler (1986, 2002), Richardson (1994) and Agresti (2013, Chapter 17).

[1] To refrain from including too big a list of books aligned with the analysis of contingency tables, this list excludes books whose topics are primarily concerned with specific techniques that are often applied to analyse them including, but limited to, log-linear and association models, measures of association, quantification theory, correspondence analysis and ecological inference.

The 42 books just mentioned cover a very broad range of topics including the development of correlation, the chi-squared statistic, and other measures of association. Their development was dominated by those from the UK in the first half of the 20th century with notable contributions by Sir Francis Galton (1822 – 1911), Karl Pearson (1857 – 1936), George Udny Yule (1871 – 1951), Sir Ronald. A. Fisher (1890 – 1962), John B. S. Haldane (1892 – 1964), Maurice S. Bartlett (1910 – 2002) and Robin L. Plackett (1920 – 2009) while the second half of the century saw those from the USA making key developments including Leo A. Goodman (1928 – 2020), Stephen E. Fienberg (1942 – 2016), Shelby J. Haberman and Alan Agresti. More recently, the visualisation of the association that exists between categorical variables has gained widespread attention. This is especially so in the correspondence analysis literature where historical and bibliographical contributions can be found in de Leeuw (1983), Friendly (2002, 2006), Beh (2004), Armatte (2008), Beh & Lombardo (2012, 2014, 2019) and Cuadras & Greenacre (2022). Friendly (2006) and Meyer, Zeileis & Hornik (2008) also provide further reviews of the history of data visualisation and its use for analysing contingency tables.

2 A Brief Overview of Tornado Predictions

Tornado research has been an ongoing area of interest in the meteorological disciplines for over a century and the observation of tornadoes has been made for much longer. Bradford (1999) points out that tornadoes have been described in the Bible, Aristotle's *Meterologica* and Pliny the Elder's *Naturalis Historia*. Bradford (1999) also says that it is likely that the first recorded tornado in the newly established American colonies was reported by the John Winthrop, former Governor of Massachusetts (1630 – 1634). In a diary entry on July 5, 1643, Winthrop includes a description of a violent wind gust that blew down trees and "lifted up their meeting house at Newbury, people being in it". While tornadoes do blow down trees and lift buildings, it has not been confirmed that this wind gust was a result of a tornado.

TABLE NO. I.
Tornado Predictions and Verifications.

MONTH.	Predictions for	Total number.	Number of predictions "favorable for tornadoes."	Fully verified.	Number of predictions "unfavorable for tornadoes."	Fully verified.	Total number made.	Total number fully verified.
March	8 hours	771	43	6	728	721	771	727
April	8 hours	934	25	11	909	906	934	917
May	8 hours	558	10	8	548	542	558	550
May	16 hours	549	22	3	518	511	549	514

FIGURE 1. Finley's (1884b, p. 86) data cross-classifying the number of tornadoes observed with the number of tornadoes predicted in the 18 US districts. March – May, 1884.

More formal and accurate records of storm activity were being taken after the public demanded the US government investigate their occurrence. Galway (1992) describes that with the invention of the telegraph in 1833 and its availability to the public in April 1845, this allowed for the development of a quick and reliable communication system for weather services in the US and across the world. The Smithsonian Institute formed a system of observations in 1847 which saw volunteers' man 150 stations that took part in the first year of weather observation. This grew to about 500 stations by 1860. However, the US Civil War reduced those numbers but in 1873 the Director of the Smithsonian petitioned the chief signal officer of the US Army and head of the Signal Service to resume funding the stations. So, on New Years Day of 1874 the Signal Service was responsible for the activities of the stations although "little effort was made to compile a data bank for research purposes" (Galway, 1992, p. 565). Although, by 1882, tornado investigations began in the US with about 800 observers put in charge of reporting when tornadoes and other meteorological events occurred. It was in that year that Seargent John Park Finley (1854 – 1943) was put in charge of a project to study tornadoes (Galway, 1992, p. 565). Not long after his appointment, Finley (1884a, p. 767) stated in the June 20 issue of *Science* that:

> "In the study of tornadoes it has become necessary to undertake something more than a simple record of their occurrence, or an occasional investigation of those that are attended with unusual destruction to life and property. A practical knowledge of the nature of these destructive storms is a matter of the utmost importance to the inhabitants of certain sections of the country; and not least among the objects at by the chief signal-officer, in directing the continuance of tornado investigation, is to allay any needless anxiety or fear on the part of those people living in the regions most frequented by these storms."

He then described a review of seven points that outline his observations on the characteristics of tornadoes. Although it would be in the following month that Finley would gain acclaim and criticism. It was in July of 1884 that Finley published an article in the *American Meteorological Journal* that included a summary of his prediction of tornadoes and their observation across 18 districts of the US (Finley, 1884b). These observations and predictions are given here as Figure 1. So it is with this article that our discussion starts as we describe the feedback that it received during 1884 and 1885. In particular, we shall be discussing the responses made by Gilbert (1884), Peirce (1884) and Doolittle (1885) who each commented on Finley's data and his index by putting forward alternative indices; such a discussion has been aptly referred to as the "Finley affair" by Murphy (1996). These discussions involve data that is best summarised in the form of a 2×2 contingency table and the development of their indices can be regarded as the pre-history of some of the key moments in the development of categorical data analysis that were made by pioneers including Sir Francis Galton (1822 – 1911), Karl

Pearson (1857 – 1936) and G. Udny Yule (1871 – 1951). The responses to Finley's (1884b) paper made by Grobe Karl Gilbert (1843 – 1918), Charles Sanders Peirce (1839 – 1914) and Myrick Hascell Doolittle (1830 – 1913) adopt a standard set of notation (for 1884 and 1885) although we shall be using the notation in Table 1 to describe the key features that came from these discussions.

We do point out that Armistead (2016) also discussed the contributions of Finley, Gilbert, Peirce, and Doolittle but did so by describing the pre-history of a probability measure that "mirrored" Bayes' theorem. However, this paper gives context of the contributions of Finley, Gilbert, Peirce and Doolittle in relation to the contributions of the early pioneers of contingency table analysis. We also recognise that Goodman & Kruskal (1959, Section 3.1) gave an interesting account of the contributions made by Finley, Gilbert and others and who viewed their contributions through the lens of one who is analysing a 2×2 contingency table. We shall provide further historical notes and comments and discuss properties of each of the indices that were not previously discussed.

Table 1. Notation of a 2×2 contingency table

	Occurrence		
Prediction	Tornado	Not Tornado	Total
Tornado	n_{11}	n_{12}	$n_{1\bullet}$
Not Tornado	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

3 Finley's Tornado Predictions

3.1 Finley's Data

In Section 2 we discussed Finley's concern that studies of tornadoes should involve more than just a "simple record of their occurrence". There is plenty of meteorology literature that details Finley's life and contributions to tornado research and their prediction; see, for example Galway (1985, 1992), Murphy (1996), Bradford (1999), Grice *et al*. (1999) and Hogan *et al.* (2009). We shall leave it to the reader to pursue these accounts at their leisure and restrict our introduction of Finley to the nature of his data and his *index of verification*, an index he proposed for assessing the accuracy of his tornado predictions.

Finley's observations and predictions of his data are given in Figure 1. His first data set comes from the predictions he made of a tornado occurring (or not) and the observations recorded on March 10, 1884, across 18 districts of the US. Finley's data are based on the observations and predictions he made that lie east of 105° longitude with "tornado alley" being on the western limit of this region. He started his predictions of whether a tornado would occur or not during the eight-hour period that day starting at 7am (Washington time). A second set of predictions was then made at 3pm for the eight-hour period until 11pm. Further predictions of whether a tornado would be observed were made in April and twice in May. The data from Finley's observations and predictions in April 1884 appear as a 2×2 contingency table in Table 2 where 934 predictions were made. This data was also examined by Goodman & Kruskal (1959, p. 128).

Table 2. Finley's tornado data of April 1884.

	Occurrence		
Prediction	Tornado	Not Tornado	Total
Tornado	11	14	25
Not Tornado	3	906	909
Total	14	920	934

3.2 Finley's Index

Finley's April results suggest that he correctly predicted 11 tornadoes occurring and 906 tornadoes not occurring during the 8-hour period. He therefore calculated his *index of verification*, being the probability of successfully predicting whether a tornado would be observed or not as:

$$i_{\mathrm{F}} = \frac{11 + 906}{934} = 0.9818\,.$$

This index[2] suggests that 98.18% of the forecasts that Finley made were correct, an impressively high percentage. Finley did not perform any of his calculations using any form of notation although, given the notation of Table 1, his index takes the form:

$$i_F = \frac{\mathrm{n}_{11} + \mathrm{n}_{22}}{\mathrm{n}}\,.$$

Repeating the calculation of his April index for the observations/predictions he recorded in March, May (8-hour observation) and May (10-hour observation) produces an i_{F} index of 0.9429, 0.9857 and 0.9519, respectively. By aggregating each of the cell frequencies of the four 2×2 contingency tables from the data given in Figure 1, the overall probability of Finley (1884b) successfully predicting whether a tornado would be observed or not is 0.9661.

[2] Note that the subscript "F" has been added to the index to distinguish it from measures proposed by others. The first initial surname letter of the contributing author has been subscripted for the indices that will follow.

4 Gilbert's Analysis

4.1 Re-examining Finley's Analysis

Immediately following Finley (1884b) came the first of several studies that would, in less than two years, spark a great deal of discussion on the legitimacy of his index, i_F, and propose significant improvements. The first of these was made by Gilbert (1884) two months after the publication of Finley's data, in the September issue of *The American Meteorological Journal*.

Finley's skill in being able to predict tornadoes in April 1884 with more than 98% accuracy appears astounding. However, his index was very much under the influence of the (2, 2)th cell frequency which, as Table 2 shows, accounts for $100 \times 906/934 = 97\%$ of the sample size (n = 934). This was a point of contention raised by "G" in a *Letter to the Editor* of *Science* that appeared a month after Finley published his data and index. G pointed out that:

> ". . . this remarkably high percentage of verification is largely made up, *not* of successful predictions of tornadoes, but of successful predictions of no tornadoes" (G, 1884, p. 126)

then colourfully adding that (for Finley's analysis of the March results):

> "An ignoramus in tornado studies can predict no tornadoes for a whole season, and obtain an average of fully ninety-five percent. The value of the expert work must, therefore, be measured by the excess which is obtained over the man who knows nothing of the subject." (p. 126)

The question then is *how does one verify the prediction of tornadoes NOT occurring?* especially when $n_{22} \gg n_{11}$. Adding to G's (1884) comments, Gilbert (1884) also pointed out the fallacious nature of Finley's results. Despite his concerns of Finley's analysis, Gilbert (1884) still admired Finley's attempt saying at the start of his discussion:

> "The account of tornado predictions given by Sergeant Finley . . . is of great interest to those who are sanguine of the ultimate successful forecasting of these destructive storms. In my judgment it shows encouraging progress, and if I take the occasion to point out a fallacy in his discussion, the reader must not understand that I undervalue the general results of his investigations." (p. 166)

Unlike G (1884), Gilbert (1884) then discusses an alternative index to i_F as we shall now describe.

4.2 Gilbert's Index

Gilbert (1884) argued that predicting the number of tornadoes to occur should be based not just on the "favourable" predictions but also on the "unfavourable" predictions – since the final prediction should be determined by how successful a prediction is (like Finley, 1884b, did) AND how unsuccessful the prediction is. So, Gilbert defined:

- the number of "favourable" predictions to be those that occurred and those that did not occur, being is n_{11} and $(n_{1\bullet} - n_{11})$ respectively. Therefore, the total number of favourable predictions is $n_{11} + (n_{1\bullet} - n_{11}) = n_{1\bullet}$.
- the number of "unfavourable" predictions to be the number of tornadoes that occurred but were not predicted, this quantity being $n_{\bullet 1} - n_{11}$.

Therefore, Gilbert's (1884, eq. (1)) probability of successfully predicting whether a tornado would be observed or not is:

$$v = \frac{n_{11}}{n_{1\bullet} + (n_{\bullet 1} - n_{11})}$$

and he referred to it as the *ratio of verification*. Gilbert's index can also be expressed alternatively, and equivalently, by:

$$v = \frac{n_{11}}{n - n_{22}} = \frac{n_{11}}{n_{11} + n_{12} + n_{21}}.$$

Gilbert's original definition of his index is still being used in the meteorological community and is referred to as the *Critical Success Index* (CSI) or *Threat Score*; see, for example, Schaeffer (1986, 1990) and Hogan *et al.* (2009). It should be noted that after Gilbert (1884) derived his index, v, it was later independently derived across a range of other disciplines. For example, it is equivalent to the *similarity coefficient* discussed by Sneath (1957, p. 13) who was concerned with bacterial classification. It is also equivalent to the coefficient discussed by Jaccard (1912) who studied the distribution of flora along the French/Swiss Alpine region.

Interestingly, unlike Finley's index, Gilbert's index does not include n_{22}. When describing the various indices that can be obtained from a 2×2 contingency table (at least for ecological purposes), Janson & Vegelius (1981) suggest that all indices should be independent of n_{22}, the number of "negative matches", noting that if n_{22} is not ignored it:

> ". . . would tend to give a high value [of the index], indicating a high degree of coexistence between very rare species, even if they are seldom found together" (p. 371).

This is certainly the case for Finley's results and resonates with G's (1884) comments. Gilbert was also aware of this saying of Finley's observations:

> "The occurrence of tornadoes in any given one of the districts indicated by him, is highly exceptional; their non-occurrence is the rule; and this consideration is

> overlooked when the predictions of occurrence and non-occurrence are classed together as of equal difficulty." (p. 166)

Therefore, calculating Gilbert's (1884) *ratio of verification* for Table 2, the probability of successfully predicting whether a tornado will occur or not is:

$$v = \frac{11}{14 + 25 - 11} = 0.3929$$

which certainly appears far more reasonable than what Finely's index of 0.9818 would suggest. Similarly, the value of v for the March, May (8-hours) and May (10-hours) observations are 0.1200, 0.5000 and 0.1034, respectively. Aggregating the four 2×2 tables gives an overall ratio of verification of 0.2276. Without knowing the more about the behaviour of v for each data set its value cannot clearly interpreted. Therefore, we shall discuss how reasonable each of the indices, described in this paper are, including i_F and v, in Section 7.

4.3 Gilbert's Revised Index and "e"

Gilbert (1884) was not completely satisfied with his v. He realised that one aspect of his analysis that was missing was how good the predictions were compared to whether a prediction had been made "fortuitously". Gilbert provided a clear explanation of the ratio of what is observed to what is expected to occur by "coincidence" saying:

> "It is to be observed, however, that the ratio of verification falls far short of a just measure of success in scientific forecasting, for with the same skill in inference this ratio may be larger or smaller according as the phenomena foretold are normally frequent or rare." (p. 168)

Therefore, he adjusted his v by first noting that (keeping to his original spelling and using the notation of Table 2):

> "If the forcaster were to make his $\mathrm{n}_{1\bullet}$ predictions at random it is probable that a certain number, e, of predictions would fortuitously coincide with occurrences. Making his predictions by the aid of inference, the number of coincidences is n_{11}. $\mathrm{n}_{11} - e$ coincidences are thus the product of his skill in inference, and $\mathrm{n}_{11} - e$ may be regarded as a measure of his success in inference, in precisely the same sense in which n_{11} has been regarded above as a measure of verification." (p. 168)

In fact, such a comment reflects perfectly with G's (1884, p. 126) statement given in Section 4.1:

> ". . . The value of the expert work must, therefore, be measured by the excess which is obtained over the man who knows nothing of the subject."

It seems then that this *excess* is exactly what e assesses and so it may be conjectured that "G" is in fact "Gilbert", although there appears nothing in the literature to substantiate this claim.

Gilbert (1884) then defined a second equation, i, which he referred to as the *ratio of success inference*. His ratio i, expressed as equation (2) in his paper, is defined using the notation of Table 2 as:

$$i_G = \frac{n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{n}}{(n_{\bullet 1} + n_{1\bullet} - n_{11}) - \frac{n_{1\bullet}n_{\bullet 1}}{n}}$$

where his e coincides with $n_{1\bullet}n_{\bullet 1}/n$. This index is, in principle, the same as v but the number of tornadoes occurring by "coincidence" has been deducted from the numerator and the denominator of v. Therefore, the probability of correctly predicting whether a tornado will occur, or not, taking into consideration any that occur by chance, is revised from 0.3929 to:

$$i_G = \frac{11 - \frac{25 \cdot 14}{934}}{(14 + 25 - 11) - \frac{25 \cdot 14}{934}} = 0.3846\,.$$

Note that Gilbert (1884, p. 168) referred to $n_{11} - n_{1\bullet}n_{\bullet 1}/n$ as the measure of *success in inference* which would later be termed a *contingency* by Pearson (1904, p. 5); a term that serves as the etymology of *contingency table*. Although, Pearson (1904) twice used the phrase *pure contingency table* rather than *contingency table*, the first time on page 33 when he describes a table of counts formed from the cross-classification of the occupation of 775 pairs of fathers and sons (Pearson, 1904, p. 33) and the second time on the page 34 where he discusses the "*theoretical* importance" of the "general conception of contingency". Gilbert (1884, p. 169) describes further his justification for subtracting the quantity $n_{1\bullet}n_{\bullet 1}/n$ from the numerator and denominator of v as follows:

> ". . . in case of random prognostication, the ratio of the fortuitous coincidence (e) to the number of predictions [$n_{\bullet 1}$] is equal to the ratio of the occurrences [$n_{1\bullet}$] to the total of cases – occurrences and non-occurrences [n]
>
> $$\frac{e}{n_{1\bullet}} = \frac{n_{\bullet 1}}{n} \quad \text{or} \quad e = \frac{n_{\bullet 1}n_{1\bullet}}{n}.\text{"}$$

What should be immediately clear here is that while Gilbert's interpretation of e is that it is the number of correctly predicted tornadoes if the predictions were made by "chance" or "coincidence", it is therefore the expected frequency of the (1,1)'th cell of a contingency table if the null hypothesis is that there is no association between the two categorical variables. Goodman & Kruskal (1959, p. 129) note that Gilbert's i_G is zero when the observed proportion of cell counts in the (1, 1)th cell is equivalent to what is expected if the variables are not associated but they make no further comment on the origins e. Since the numerator of i_G is just the "contingency" of the (1, 1)th cell under independence then $i_G = 0$ and this is consistent with most measures of associations used today. It isn't obvious yet whether $i_G = 0$ differs practically or not to the observed value of $i_G = 0.3846$ however we shall be assessing this difference by determining the various features of i_G in Section 7.6.

4.4 Galton and the Expected Cell Count

In 1892 Galton published a book titled *Finger Prints* that is considered to be the genesis of finger print analysis performed today; see, for example, the discussions made by Stigler (1995) and Gillham (2001). Stigler (2002) also commences his discussion of the pre-history of the $\mathrm{I} \times \mathrm{J}$ contingency table (where $\mathrm{I}, \mathrm{J} > 2$) by referring to Galton's 1892 book.

By studying a sample of 105 sets of twin brothers, or *couplets*, Galton was interested in determining the expected number of pairs with distinct fingerprint characteristics (those being arches, whorls and loops) in their right forefinger; Figure 2 gives the data Galton analysed, appearing as a 3×3 contingency table. On page 174 of his book, Galton set one set of twin brothers totals to A and the other twin brothers to B and said:

> "The question, then, was how far calculations from the above data [Figure 2] would correspond to the [the observed random couplets]. The answer is that is does so admirably. Multiply each of the . . . A totals into each of the . . . B totals, and after dividing each result by [n]"

Galton (1892, pp. 175 – 176) goes on to say of Table 2:

> "The squares that run diagonally from the top at the left, to the bottom at the right, contain the double events, and it is with these that we are now concerned. Are entries in those squares larger or not than the randoms . . . The values of 10x19, 68x61, 27x25, all divided by 105?"

Galton referred to his expected cell frequencies as *calculated random couplets* but their calculation now commonly appears in its simplest form as:

$$\text{Expected cell frequency} = \frac{\text{row total} \times \text{column total}}{\text{sample size}}.$$

Therefore, Galton's expected value of a cell frequency is just Gilbert's (1884) e. While Galton was interested in a general expression for this expected value he was only interested in comparing the observed cell counts with their expected value along the diagonal of his 3x3 table. Note that Gilbert (1884) was only concerned with the (1, 1)th cell of a 2×2 contingency table because he was interested only in verifying the prediction of tornadoes and not in verifying that tornadoes did not occur. If he had, then Gilbert's analysis of Finley's data (when viewed as a 2 × 2 contingency table) is identical in nature to Galton's interest in his fingerprint data.

It therefore seems fair to assign some credit for the derivation and justification of an expected cell frequency of a contingency table to Gilbert (1884). Although, some may argue against this since Finley, and Gilbert, did not present or even analyse the data in Figure 1 as a 2 × 2 contingency table. It is also recognised that Galton's impact on Pearson's work is profound and is well documented; see, for example, Yule & Filon (1936), Haldane (1957) and Gillham (2001). One can therefore understand why, during the emergence of statistical thinking that was taking place in England at the turn of the 20th century, that Gilbert's (1884) contribution to e would be overshadowed by Galton's (1892) contribution. In fact, prior to Pearson (1904), Galton's expression of the expected cell frequency was also described for a 2 × 2 contingency table by Yule (1900, §16; 1903 eq. (7)). More will be said on the apparent lack of acknowledgement of Finley's and Gilbert's (and others) work by Galton, Pearson and their successors in Section 9.2.

TABLE XXII.

Observed Fraternal Couplets.

B children.	A children.			Totals in B children.
	Arches.	Loops.	Whorls.	
Arches . . .	5	12	2	*19*
Loops . . .	4	42	15	*61*
Whorls . . .	1	14	10	*25*
Totals in A children	*10*	*68*	*27*	105

FIGURE 2. Galton's (1892, p. 175) original data cross-classifying fingerprint types between two male fraternal twins, or *couplets*

5 Peirce's Analysis

Following on from Gilbert's (1884) analysis of Finley's data (Figure 1) is the contribution of Charles Sanders Peirce. After defining notation that is very similar in form to the notation used by Yule (1900, 1903) for a 2 × 2 contingency table, Peirce (1884) says:

> "Then the problem is . . . to assign a numerical measure to the success or science of the method by which the answers have been produced. Mr G.K. Gilbert . . . has recently proposed a formula for this purpose and I desire to offer another."

Peirce's description of his index starts colourfully by discussing his interest in studying the difference between the observations of an "infallible witness" and the observations of "an utterly ignorant person" – the latter can be more diplomatically described as observations that would arise purely by chance. Peirce (1884) would propose his amendment of Gilbert's, and Finley's, index by also denoting it also by i and defining it as the:

"... proportion of questions put to the infallible witness" (p.453)

It is implied here that the questions put to this "witness" are correctly answered, although his derivation of his index shows a slightly different interpretation. Framing Peirce's (1884) description in terms of Finley's (1884b) data, Peirce's wanted to determine the difference between correctly and incorrectly predicting that a tornado will occur. Using such terms, his i (we shall denote it as i_P) is therefore the proportion of tornadoes correctly predicted and he referred to it as the *measure of the science of the method*. He also defined j (here, j_P) to be the

> "... proportion of questions which the ignorant witness answers in the first way" (p. 453)

Here, "first way" refers to the witness incorrectly observing the occurrence (or not) of a tornado. Peirce goes on to determine i_P and j_P by solving the following four equations:

$$\mathrm{n}_{11} = i_P \mathrm{n}_{\bullet 1} + (1 - i_P) j_P \mathrm{n}_{\bullet 1}$$

$$\mathrm{n}_{12} = (1 - i_P) j_P \mathrm{n}_{\bullet 2}$$

$$\mathrm{n}_{21} = (1 - i_P)(1 - j_P) \mathrm{n}_{\bullet 1}$$

$$\mathrm{n}_{22} = i_P \mathrm{n}_{\bullet 2} + (1 - i_P)(1 - j_P) \mathrm{n}_{\bullet 2} \, .$$

The first two of these equations can be expressed as the relative cell frequencies of the first row (ie the prediction of a tornado) such that:

$$\frac{\mathrm{n}_{11}}{\mathrm{n}_{\bullet 1}} = i_P + (1 - i_P) j_P$$

$$\frac{\mathrm{n}_{12}}{\mathrm{n}_{\bullet 2}} = (1 - i_P) j_P$$

and yields the solution to his index:

$$i_P = \frac{\mathrm{n}_{11}}{\mathrm{n}_{\bullet 1}} - \frac{\mathrm{n}_{12}}{\mathrm{n}_{\bullet 2}} .$$

This is the first solution Peirce (1884) gave to his index and is the difference between the proportion of observed tornadoes that were predicted and the proportion of tornadoes that were not observed but were predicted. Armistead (2016) describes i_P as being the difference between the "true positive fraction" (that is, correctly identifying the occurrence of a tornado) and the "false positive fraction" (that is, incorrectly predicting the occurrence of a tornado). Thus, while we have shown in Section 4.3 that Gilbert's index for Table 2 is $i_G = 0.3846$, Peirce's (1884) *measure of the science of the method* of this data is:

$$i_P = \frac{11}{14} - \frac{14}{920} = 0.7705\,.$$

Therefore, Peirce's index appears to show a stronger positive "link" between the prediction of tornadoes and what was observed when compared with Gilbert's index.

While Peirce did not give a statement for j_P, the second of his four equations yield:

$$j_P = \frac{\mathrm{n}_{12}/\mathrm{n}_{\bullet 2}}{\mathrm{n}_{21}/\mathrm{n}_{\bullet 1} + \mathrm{n}_{12}/\mathrm{n}_{\bullet 2}}$$

so that it is the probability of a predicted tornado not occurring given that the predictions made were all incorrect.

The last two of Peirce's four equations can be expressed as relative cell frequencies of the second row (ie the prediction of a tornado not occurring) so that:

$$\frac{\mathrm{n}_{21}}{\mathrm{n}_{\bullet 1}} = (1 - i_P)(1 - j_P)$$

$$\frac{\mathrm{n}_{22}}{\mathrm{n}_{\bullet 2}} = i_P + (1 - i_P)(1 - j_P)\,.$$

Hence, while Peirce did not show this, i_P can also be defined as:

$$i_P = \frac{\mathrm{n}_{22}}{\mathrm{n}_{\bullet 2}} - \frac{\mathrm{n}_{21}}{\mathrm{n}_{\bullet 1}}$$

which is the difference between the proportion of tornadoes that did not occur but were (correctly) not predicted and the proportion of tornadoes that were observed but were not predicted. This definition of the index also gives $i_P = 0.7705$ for Table 2.

Solving all four equations yields the solution:

$$i_P = \frac{\mathrm{n}_{11}\mathrm{n}_{22} - \mathrm{n}_{12}\mathrm{n}_{21}}{\mathrm{n}_{\bullet 1}\mathrm{n}_{\bullet 2}}$$

which is identical Youden's (1950, p. 33) J-index. Youden's motivation for deriving his index was the same as Peirce's but their contexts were very different; Peirce was concerned with correctly identifying that a predicted tornado occurred while Youden was concerned with correctly identifying that someone was diagnosed to have a disease. This definition of the index also gives $i_P = 0.7705$ for Table 2.

Peirce (1884) goes on to compare his index with Gilbert's which he shows to be equivalent to:

$$i_G = 2\frac{n_{11}n_{22} - n_{12}n_{21}}{n^2 - n_{11}^2 + n_{12}^2 + n_{21}^2 - n_{22}^2}$$

so that, for Table 2:

$$i_G = 2\frac{11 \cdot 920 - 14 \cdot 3}{934^2 - 11^2 + 14^2 + 3^2 - 906^2} = 0.3846$$

confirming the calculation of Gilbert's index in Section 4.3. Peirce then says that he has also derived extensions of this index for categorical data consisting of more than two categories:

> "If the questions should present more than two alternatives, it would be necessary to assign relative values or measures to the different kinds of mistakes that might be made. I have a solution for this case" (p. 454)

although he does not present his solution, nor, unfortunately, does he direct the reader to where such a solution can be found.

Rovine & Anderson (2004) also provide a derivation of Peirce's (1884) i_P and notes that it has the hallmarks of a coefficient of association described by Yule (1900). These being that $i_P = 0$ when the two variables are independent – since $n_{11}n_{22} = n_{12}n_{21}$; a feature of independence described by Yule (1900, §19). Also, $i_P = 1$ when there is perfect positive association and $i_P = -1$ when there is perfect negative association; we shall discuss these features further in Section 7.7. In fact, in the case where $n_{1\bullet} = n_{\bullet 1}$ and $n_{2\bullet} = n_{\bullet 2}$ then Peirce's index is equivalent to the square root of Pearson's (1904, eq. (xxviii)) mean square contingency and so is also a correlation coefficient.

6 Doolittle's (1885) Analysis

6.1 Deriving his Index

Following on from Finley's (1884b) study of his tornado data and Gilbert's (1884) response to Finley's work, further emendations were undertaken by Doolittle (1885). Doolittle was an excellent mathematician and famous for the contributions he made to matrix algebra, especially for being the first to propose a more efficient method of performing Gaussian elimination (Doolittle, 1878); see, for example, Dwyer (1941) and Grcar (2011). He also demonstrated his talents a few years after this when, in 1884, he turned his attention to studying Finley's data and amending Gilbert's index. Doolittle adopted the same notation used by Gilbert (1884) and introduced his paper by saying:

> "Mr G. K. Gilbert has published . . . a method of estimating the ratio of skill in predictions of occurrences and non-occurrences of a simple event." (p. 122)

While Doolittle refers to "a simple event" he does explain his development of his index in general terms before analysing Finley's tornado data. Doolittle (1885, p. 123) then notes that the probability of "success is proportional" to $n_{11}/n_{1\bullet}$ and $n_{11}/n_{\bullet 1}$; here he is talking about the proportion of successfully predicted tornadoes to occur AND, based on the predictions that are made, the proportion of tornadoes that occurred, respectively. Therefore, Doolittle (1885) defined the index as:

$$s_D = \frac{n_{11}}{n_{\bullet 1}} \cdot \frac{n_{11}}{n_{1\bullet}}$$

which he refers to as the proportion of "successful" predictions. Therefore, for Table 2, this index is equal to:

$$s_D = \frac{11}{14} \cdot \frac{11}{25} = 0.3457\,.$$

Doolittle was aware that s_D does not accommodate for the possibility of observations happening by "chance" but does say that:

> "The fraction [$n_{\bullet 1}/n$] represents the ratio of random success and therefore [$n_{\bullet 1}n_{1\bullet}/n$] verifications out of [$n_{1\bullet}$] predictions are to be ascribed to chance and must be subtracted throughout." (p. 123)

There are two things to note here. Firstly, he notes that of the $n_{1\bullet}$ predicted tornadoes, there are $n_{\bullet 1}n_{1\bullet}/n$ of them that occur by "chance". This is precisely Gilbert's e and is the expected number of correctly predicted tornadoes to be observed if the predicted number of tornadoes and the observed number of tornadoes was completely independent; that is, the predictions were no different to "chance". Secondly, by saying "throughout" Doolittle is referring to each of the terms on the numerator and denominator of his index, i_D – something he does not immediately do. Rather, Doolittle first describes $n_{\bullet 1}(1 - n_{1\bullet}/n)$ and $n_{1\bullet}(1 - n_{\bullet 1}/n)$ to be:

"... fields which chance leaves for science to conquer"

and is now known to be the variance of the number of tornadoes to be observed, and predicted, respectively, when n_{11} is assumed to follow a binomial distribution where the 2×2 table has fixed and known marginal frequencies. Peirce also says of $n_{11} - n_{1\bullet}n_{\bullet 1}/n$ that it is:

"... the portion of each which science does conquer".

It is on the next line of his discussion that Doolittle subtracted $n_{\bullet 1}n_{1\bullet}/n$ "throughout" and in doing so follows the same tact that Gilbert used when he derived his index, i_G. That is, Doolittle amended his original index so that it is of the form:

$$i_D = \frac{n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{n}}{n_{\bullet 1} - \frac{n_{1\bullet}n_{\bullet 1}}{n}} \cdot \frac{n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{n}}{n_{1\bullet} - \frac{n_{1\bullet}n_{\bullet 1}}{n}}$$

which he referred to as the *degree of logical connection* between the observed number of tornadoes and the predicted number of tornadoes. It is at which point that he also gives an alternative form of this revised index showing that i_{D} is also equivalent to:

$$i_{\mathrm{D}} = \frac{(\mathrm{n}\mathrm{n}_{11} - \mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1})^2}{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}(\mathrm{n} - \mathrm{n}_{1\bullet})(\mathrm{n} - \mathrm{n}_{\bullet 1})}$$

which can also be written as:

$$i_{\mathrm{D}} = \mathrm{n}^2 \frac{\left(\mathrm{n}_{11} - \frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}}\right)^2}{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 2}}.$$

Therefore, Doolittle's revision of Gilbert's index for Table 2 gives an index value that is comparable to his own $s_D = 0.3457$ and Gilbert's index (of $i_G = 0.3929$) where:

$$i_{\mathrm{D}} = \frac{(934 \cdot 11 - 25 \cdot 14)^2}{25 \cdot 14 \cdot (934 - 25)(934 - 14)} = 0.3365$$

but differs substantially to Peirce's index of $i_P = 0.7705$.

Doolittle then proceeds to derive this same index a second way. This time noting that (keeping to his original spelling):

> "Since the skillful predictions are mingled indistinguishably with all unskilled ones, and are vitiated accordingly, the value of the *vitiated* probability of the skillfil prediction of any single occurrence may be represented by the product

$$i_{\mathrm{D}} = \left(\frac{\mathrm{n}_{11}}{\mathrm{n}_{\bullet 1}} - \frac{\mathrm{n}_{1\bullet} - \mathrm{n}_{11}}{\mathrm{n} - \mathrm{n}_{\bullet 1}}\right)\left(\frac{\mathrm{n}_{11}}{\mathrm{n}_{1\bullet}} - \frac{\mathrm{n}_{\bullet 1} - \mathrm{n}_{11}}{\mathrm{n} - \mathrm{n}_{1\bullet}}\right) = \frac{(\mathrm{n}\mathrm{n}_{11} - \mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1})^2}{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}(\mathrm{n} - \mathrm{n}_{1\bullet})(\mathrm{n} - \mathrm{n}_{\bullet 1})}\ ."$$

By saying *vitiated* Doolittle concedes that determining the probability of making a successful prediction is "spoiled" by any randomness that may exist in the process of calculating a successful outcome. Therefore, he deals with this "spoiled" prediction by removing it from the observed number of successful predictions, just as Gilbert did. This can be seen by rewriting i_{D} in a slightly different, but equivalent, way:

$$i_{\mathrm{D}} = \left(\frac{\mathrm{n}_{11}}{\mathrm{n}_{\bullet 1}} - \frac{\mathrm{n}_{12}}{\mathrm{n}_{\bullet 2}}\right)\left(\frac{\mathrm{n}_{11}}{\mathrm{n}_{1\bullet}} - \frac{\mathrm{n}_{21}}{\mathrm{n}_{2\bullet}}\right).$$

We can see here that this index removes from the two probabilities of "success" of predictions and observations their associated probabilities of "failure".

6.2 Doolittle and the Concept of "Contingency"

Recall in Section 6.1 that Doolittle's derivation of his index i_{D} involved calculating the difference $\mathrm{n}_{11} - \mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}/\mathrm{n}$. This measure is often attributed to Pearson (1904) who considers a more general difference, $n_{uv} - n_{u\bullet}n_{\bullet v}/n$, it being for the (u, v)th cell of a $s \times t$ contingency table where $s > 2$ and $t > 2$. Thus, while Doolittle (1885), like Finley (1884b) and Gilbert (1884) before him, were concerned with the how to verify tornado predictions summarised in the form of a 2×2 contingency table, Pearson (1904) was concerned with larger sized tables. Pearson (1904, p. 5) states that his difference $n_{uv} - n_{u\bullet}n_{\bullet v}/n$ is the:

> ". . . the deviation from independent probability in the occurrence of the groups A_u, B_v".

Here Pearson defined A_u and B_v to be the u'th row and v'th column of his contingency table. On the line following this definition, Pearson (1904, p. 5) goes on to say of the difference:

> "I term any measure of the total deviation of the classification from independent probability a measure of its *contingency*. Clearly the greater the contingency, the greater must be the amount of association or of correlation between the two [categories], for such association or correlation is solely a measure from another standpoint of the degree of deviation from independence of occurrence".

While Doolittle (1885) did not discuss his index in terms of "association" or "correlation" (terms that would not come into the statistics vernacular for at least a decade after Doolittle's paper), his use of "chance" in his description of e and the phrase "science does conquer" characterise these two terms and so they can both be seen to view the difference in very similar ways. For historical perspective, David (1995) notes that "correlation" was used by Galton (1888) in his description of the relationship between the length of one's arm and their leg, while David (1998) notes that the (statistical) use of "association" was first used by Yule (1900) in his discussion of categorical data. However, in December 1888, Galton published a paper in the *Proceedings of the Royal Society of London* where the first sentence describes that the origins of "correlation" can be traced back, at least conceptually if not quantitatively, to early work in biology:

> " 'Co-relation or correlation of structure' is a phrase much used in biology, and not least that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree". (Galton, 1888, p. 135)

The reader is invited to read Stigler (1989) who gave an interesting discussion of Galton's account of the conception of correlation. Interestingly, Doolittle followed up his 1885 paper in 1888 where he presented a range of quantities for a 2×2 contingency table. This paper, which Doolittle presented to members of the *Philosophical Society of Washington (Mathematical Section)* on February 16, 1887, describes each quantity as being the "extent of association" for the cell of the table, or for the table itself. While Doolittle did not provide an extensive discussion on the philosophical meaning and interpretation of "association" like that espoused by Galton, Pearson and Yule, his usage of the term is very much consistent with their interpretation. The interested reader is invited to read this paper which is simple titled "*Association Ratios*".

6.3 Doolittle and the Mean Square Contingency

One may note that Peirce's s_D is equivalent to the (1, 1)th element of the partition of Pearson's (1904, eq. (xxviii)) mean square contingency of the 2×2 table:

$$\phi^2 = \frac{X^2}{n} = \sum_{j=1}^{2}\sum_{i=1}^{2}\left(\frac{n_{ij}}{n_{i\bullet}} \cdot \frac{n_{ij}}{n_{\bullet j}}\right) - 1 \, .$$

where X^2 is Pearson's chi-squared statistic. Therefore, there is a thread of commonality as well as substantial differences with the way Doolittle viewed his measure of "success" and how Pearson derives his mean square contingency for a contingency table. Doolittle was only interested in the (1, 1)'th element of a 2×2 contingency table while Pearson was concerned with all elements of a larger sized contingency table. This difference comes about by the nature of how Doolittle and Pearson viewed their analysis of the contingency table. Like Gilbert, Doolittle was only concerned with the verification of the tornadoes that were predicted to occur and so confined his attention to the (1, 1)th cell frequency. On the other hand, Pearson was more interested in general measures of association and so was concerned with all ALL frequencies. If Doolittle had considered determining his index for all four elements, then perhaps he would have simply summed his terms (being the simplest of operations) resulting in an emended version of s_D:

$$\tilde{s}_D = \sum_{j=1}^{2}\sum_{i=1}^{2}\left(\frac{n_{ij}}{n_{i\bullet}} \cdot \frac{n_{ij}}{n_{\bullet j}}\right)$$

which is equivalent to $\phi^2 + 1$ for a 2×2 contingency table. Therefore, predicting a tornado purely by "chance" would mean that, $\tilde{s}_D = 1$ so that any deviation away from 1 would show that there was some merit in the prediction process.

We can establish that Doolittle's index i_D is far more closely aligned to Pearson's mean square contingency than $\tilde{s}_D$. If we examine i_D more carefully, we note that it can also be written as:

$$i_D = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}$$

which **is** Pearson's (1904, eq. (xxviii)) mean square contingency! That is $i_D = \phi^2$. Thus, it certainly appears that Doolittle proposed the famous mean square contingency nearly 20 years prior to Pearson's derivation of the statistic, although the nature of its derivation differs.

Perhaps if Doolittle was concerned with not just the *vitiated proportion* of success but also considered the *vitiated number* of successful predictions he would have been apportioned some credit to the early development of the chi-squared statistic, at least for a 2×2 contingency table. However, a key difference between their analysis of the contingency table is that, unlike Pearson, Doolittle made no attempt to gain an understanding of the distributional properties of his index.

What should also be apparent is that the link between Peirce's index, ι_{P}, and Doolittle's index, ι_{D}, is that:

$$\iota_{\mathrm{D}} = \iota_P^2 \left(\frac{\mathrm{n}_{\bullet 1}\mathrm{n}_{\bullet 2}}{\mathrm{n}_{1\bullet}\mathrm{n}_{2\bullet}}\right)$$

so that the square of Peirce's index is proportional to Pearson's chi-squared statistic of a 2×2 contingency table since:

$$\mathrm{X}^2 = \mathrm{n}\iota_{\mathrm{D}} = \mathrm{n}\left(\frac{\mathrm{n}_{\bullet 1}\mathrm{n}_{\bullet 2}}{\mathrm{n}_{1\bullet}\mathrm{n}_{2\bullet}}\right)\iota_P^2 \,.$$

That is, Peirce's index, ι_{D}, is identical to Pearson's mean square contingency! When analysing Finley's data sets in Figure 1 (except for May (8-hours)), $\mathrm{n}_{\bullet 1} \approx 2\mathrm{n}_{\bullet 1}$ and $\mathrm{n}_{2\bullet} \approx \mathrm{n}_{\bullet 2}$ so that $\mathrm{X}^2 = \mathrm{n}\iota_{\mathrm{D}} \approx 0.5\mathrm{n}\iota_P^2$. This can be verified by noting that, for Table 2, $\mathrm{X}^2 = 314.3$ (without Yate's continuity correction), $\mathrm{n}\iota_{\mathrm{D}} = 934 \cdot 0.3365 = 314.3$ and $0.5\mathrm{n}\iota_P^2 = 0.5 \cdot 934 \cdot 0.7705^2 = 277.2$; since $(\mathrm{n}_{\bullet 1}\mathrm{n}_{\bullet 2})/(\mathrm{n}_{1\bullet}\mathrm{n}_{2\bullet}) = (14 \cdot 920)/(25 \cdot 909) = 0.5668$, and not 0.5, then $\mathrm{n}((\mathrm{n}_{\bullet 1}\mathrm{n}_{\bullet 2})/(\mathrm{n}_{1\bullet}\mathrm{n}_{2\bullet}))\iota_P^2 = 314.3$, as expected.

6.4 Doolittle and Multiple Dichotomous Variables

Doolittle ended his 1885 discussion of the problem of the suitability of indices designed for verifying tornado predictions by saying:

> "It has been proposed to extend the problem so as to include more classes of events of which one must happen and only one can happen in any case." (p. 126)

Although he seems to have his doubts immediately saying:

> "It seems clear to me that no single numerical expression can be a proper solution to such a problem" (p. 126)

His doubts arise from the idea that the categories of one variable are well discriminated but at least one of the remaining variables consist of categories where such discrimination remains unclear. Thus, his concerns appear to be centred on the idea that the more variables that one considers in their analysis the more likely it is that discriminating between one option and the other remains in doubt ending his paper with:

> "No single numerical expression can properly comprehend these heterogenous results" (p. 127).

Doolittle's concerns are certainly an issue for all analysts although the passage of time has brought with it a variety of techniques for contingency table analysis that help to address this issue by assigning quantities to the categories of a variable. Many of these techniques have their genesis in the psychology, psychometric and ecology literature. They have more recently entered the statistical and allied literature under various names including reciprocal averaging (Hill, 1973, 1974) and dual scaling (Nishisato, 1980, 1994, 2007). Comprehending heterogeneous categories in this way dovetails nicely into visualisation techniques such as correspondence analysis; see, for example, Greenacre (1984, 2017) and Beh and Lombardo (2014, 2021).

7 Further Evaluations of the Indices

7.1 Some Preliminary Features of Table 2

With the various indices now derived and described, we turn our attention to delving a little deeper into some of their features. To do this we shall not just consider the magnitude of the indices at the observed (1, 1)th cell frequency, like Finley, Gilbert, Peirce and Doolittle did, but instead evaluate the indices across the full range of values that n_{11} can take. This will be done by assuming that the marginal frequencies of the 2×2 contingency table are known and fixed so that n_{11} is bounded by the Fréchet (1951) bounds:

$$L = \max(0, n_{\bullet 1} - n_{2\bullet}) < n_{11} < \min(n_{1\bullet}, n_{\bullet 1}) = U\,.$$

For Table 2 $n_{11} \in [0, 14]$. Since the expected value of the (1, 1)'th cell under independence (that is, Gilbert's e) plays an important role in the definition of some of the indices, we shall also make use of:

$$e = \frac{25 \cdot 14}{934} = 0.3747$$

which is very small in comparison to the (1, 1)th cell frequency of 11. This suggests that there are more tornadoes correctly predicted than what would be expected if the predictions were, using the terms of Gilbert, Peirce and Doolittle, made "fortuitously", by "chance" or "coincidence". However, the large sample size (relative to $n_{11} = 14$) is accounted for by the very large (2, 2)th value and helps to undervalue e thereby exacerbating the points raised by G (1884, p. 126) and Gilbert (1884, p. 166); see Section 4.1. For the March, May (8-hour observation) and May (10-hour observation) data sets, e is extremely small compared to its n_{11} and sample size, being 0.7250, 0.2509 and 0.4007, respectively. Aggregating the four data sets produces $e = 1.8195$.

Confirmation of the statistical significance of the *Prediction* and *Occurrence* variables in Table 2 can be made by performing a chi-squared test of independence and will be done without using Yate's (1934) continuity correction. To do so, Pearson's chi-squared statistic can be expressed in terms of only n_{11} and the marginal frequencies by:

$$X^2(n_{11}) = n\left(\frac{n_{11} - n_{1\bullet}n_{\bullet 1}}{n_{1\bullet}n_{2\bullet}}\right)^2\left(\frac{n_{1\bullet}n_{2\bullet}}{n_{\bullet 1}n_{\bullet 2}}\right)$$

and is a quadratic function of n_{11}. Thus, Pearson's chi-squared statistic of Table 2 is:

$$X^2(11) = 934\left(\frac{934 \cdot 11 - 25 \cdot 14}{25 \cdot 909}\right)^2\left(\frac{25 \cdot 909}{14 \cdot 920}\right) = 314.268$$

so that its a p-value that is less than 0.0001. Therefore, there is a statistically significant association between the prediction of tornadoes and the observed tornadoes. However, since $X^2(n_{11})$ is linearly related to n, the large sample size (again, relative to n_{11}) helps to inflate the value of the chi-squared statistic. To accommodate this feature, many, including Mosteller (1968) and Mirkin (2001), proposed dividing the statistic by its sample size giving Pearson's (1904) mean square contingency or, equivalently in the case of a 2×2 contingency table, Doolittle's (1885) i_D. Such a measure is also used throughout the correspondence analysis literature and is referred to as the *total inertia*.

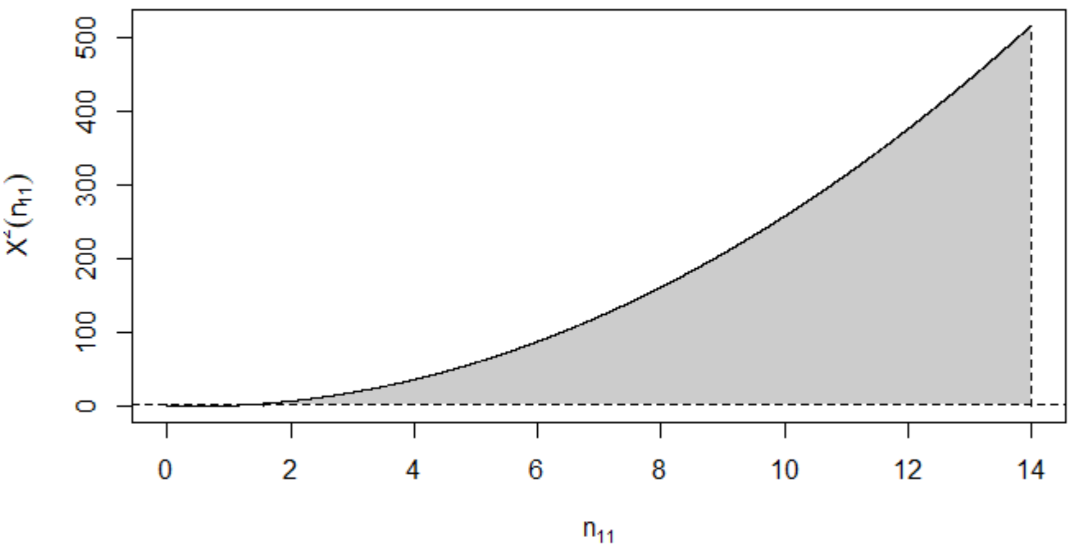

FIGURE 3. Relationship between n_{11} and $X^2(n_{11})$ for Table 2; shaded region is where a statistically significant association exists between the *Prediction* and *Occurrence* variables ($\alpha = 0.05$)

With $n_{11} \in [0, 14]$, Figure 3 shows the relationship between n_{11} and $X^2(n_{11})$ for Table 2 where the shaded region identifies where a statistically significant association exists between the variables of Table 2. This shaded region is based, in part, on the interval:

$$L_\alpha = \max\left(0, \frac{n_{1\bullet}n_{\bullet 1}}{n} - \frac{n_{1\bullet}n_{2\bullet}}{n}\sqrt{\frac{\chi^2_\alpha}{n}\frac{n_{1\bullet}n_{2\bullet}}{n_{\bullet 1}n_{\bullet 2}}}\right) < n_{11} <$$

$$\min\left(n_{1\bullet}, \frac{n_{1\bullet}n_{\bullet 1}}{n} + \frac{n_{1\bullet}n_{2\bullet}}{n}\sqrt{\frac{\chi^2_\alpha}{n}\frac{n_{1\bullet}n_{2\bullet}}{n_{\bullet 1}n_{\bullet 2}}}\right) = U_\alpha$$

where χ^2_α is the $1 - \alpha$ percentile of the chi-squared distribution with 1 degree of freedom. This is the interval of n_{11} where no statistically significant association exists and is an adaptation of the interval derived by Beh (2010) for $P_1 = n_{11}/n_{1\bullet}$. When testing the association between the *Predictions* and *Occurrence* variables of Table 2 at the $\alpha = 0.05$ level of significance, a statistically significant association exists for n_{11} lying within the interval [1.549, 14] which covers much of the interval $n_{11} \in [0, 14]$. The dominance of the (2, 2)th cell frequency plays a pivotal role in the calculation of this interval.

A comparison of the six indices plus two more (weighted versions of i_F; see Sections 7.3 and 7.4) is given in Figure 4 for Table 2 where $n_{11} \in [0, 14]$. We shall now discuss some of the key features of these indices and discuss their behaviour in terms of n_{11}. The value of each index, its bounds, and the value of index under the assumption are summarised in Table 3 for the four data sets in Figure 1.

7.2 Features of Finley's i_F

Suppose we consider Finley's index i_F. A little bit of algebra shows that it can be alternatively, and equivalently, expressed as:

$$i_F = 2\frac{n_{11}}{n} + \frac{n_{2\bullet} - n_{\bullet 1}}{n}$$

so that i_F is a linear function of n_{11}. This alternative expression tells us that, when analysing Finley's data, since $n_{2\bullet}$ is very large in comparison to $n_{1\bullet}$ and any value that n_{11} can take then $i_F \approx n_{2\bullet}/n \lesssim 1$, where "$\lesssim$" is used to mean "less than but approximately equal". For Table 2:

$$i_F = 0.002n_{11} + 0.9582 \lesssim 1\,.$$

Using the Fréchet bounds of n_{11}, the bounds of i_F are:

$$L_F = \frac{n_{2\bullet} - n_{1\bullet}}{n} < i_F < 1 + \min\left(0, \frac{2(n_{\bullet 1} - n_{1\bullet})}{n}\right) = U_F\,.$$

For Finley's data $n_{2\bullet} \gg n_{1\bullet}$, $n_{2\bullet} \lesssim n$ and $n_{1\bullet} \approx n_{\bullet 1}$; see Figure 1. Therefore,

$$L_F \lesssim i_F \lesssim U_F \lesssim 1.$$

Since the marginal frequencies of Table 2 are assumed fixed and known, i_F is bounded by $[0.9582, 0.9882]$ and this can be seen by the solid black line in Figure 3. So, the observed value of $i_F = 0.9818$ lies close to the upper bound suggesting that the predictions made by Finley are vastly better than if the predictions were made by chance. Such a conclusion would also be valid for ANY value that n_{11} can take in the interval $[0, 14]$. Therefore, since the lower and upper bound of i_F are both close to 1, irrespective of how many tornadoes that were predicted and observed, an analysis of Finley's April 1884 data using his index will mean that his predictions will ALWAYS be extremely accurate. In fact, if none of his predictions occurred, so that $n_{11} = 0$, then $i_F = 0.9582$; a value that is wholly dominated by n_{22}. If i_F is viewed in the same context as a typical measure of association with a maximum of 1, it may be interpreted as a very high value, especially when the minimum of zero is possible for a generic 2×2 contingency table. However, this is not the case. Clearly, the very large values that Finley's index take across $n_{11} \in [0, 14]$ shows how poor a quantity it is. This behaviour in the index is also observed for the March, May (8-hour observation) and May (10-hour observation) predictions and observations Finley made; the bounds of his index for these data sets are very similar to the bounds of his April records and are $i_F \in [0.9274,\ 0.9611]$, $i_F \in [0.9570,\ 0.9928]$ and $i_F \in [0.9407,\ 0.9778]$, respectively. Aggregating the cell frequencies of the four tables in Figure 1 yields the global bounds for Finley's index $i_F \in [0.9461,\ 0.9825]$.

If we were to consider Finley's index under complete independence, then:

$$\begin{aligned} i_{F|I} &= 2\frac{e}{n} + \frac{n_{2\bullet} - n_{\bullet 1}}{n} \\ &= 2\frac{n_{1\bullet} n_{\bullet 1}}{n^2} + \frac{n_{2\bullet} - n_{\bullet 1}}{n} \end{aligned}$$

where the addition of "$|I$" in the subscript indicates the index is calculated assuming there is independence between the variables of the 2×2 table. Therefore, for Table 2, the index would be:

$$i_{F|I} = 2\frac{25 \cdot 14}{934^2} + \frac{909 - 14}{934} = 0.9590$$

which lies near the lower bound of the interval $[0.9582, 0.9882]$ as does $e = 0.3747$ when compared with the range of values n_{11} can take for Table 2.

7.3 A Weighted Finley Index (Version 1)

Gilbert (1884, p. 166) remarked on the flaw induced by the large value of n_{22} when calculating Finley's index, i_F:

> "This fallacy consists in the assumption that verification of the predictions of a rare event may be classed with verifications of the predictions of frequent events, without any system of weighting."

Gilbert did not propose a weighted version of i_F but we shall provide two simple adaptations of Finley's index and examine whether there is any advantage in doing so. We concede that there may well be various other ways in which a weighted Finley index can be defined but the first one we discuss is:

$$i_F(w) = \frac{w\mathrm{n}_{11} + (1-w)\mathrm{n}_{22}}{\mathrm{n}}$$

for a given weight $w \in [0, 1]$. This index allows one to weight the predictions of a rare event (tornadoes occurring) and a frequent event (tornado not occurring) differently. It is immediately clear that if n_{11} and n_{22} were given equal weighting so that $w = 0.5$ then $i_F(0.5) = 0.5i_F$. However, regardless of the choice of w, $i_F > i_F(w)$ and this seems, on the surface, to be an acceptable property since $i_F > 0.95$ for all possible values of n_{11} given the marginal totals of Table 2. This weighted version of Finley's index can be alternatively expressed as a linear function of w by:

$$\begin{aligned} i_F(w) &= w\left(\frac{\mathrm{n}_{11} - \mathrm{n}_{22}}{\mathrm{n}}\right) + \frac{\mathrm{n}_{22}}{\mathrm{n}} \\ &= -w\left(\frac{\mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1}}{\mathrm{n}}\right) + \left(\frac{\mathrm{n}_{11} + \mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1}}{\mathrm{n}}\right). \end{aligned}$$

Since $\mathrm{n}_{11} \ll \mathrm{n}_{22}$ or, alternatively, because $\mathrm{n}_{\bullet 1} \ll \mathrm{n}_{2\bullet}$, this relationship shows that the coefficient of w is, approximately, -1 with an intercept of, approximately, $+1$ so that:

$$i_F(w) \approx -w + 1$$

and the magnitude of $i_F(w)$ is dominated far more by the choice of w than the cell frequencies of the 2×2 contingency table. For example, the relationship between $i_F(w)$ and w for Table 2 is:

$$i_F(w) = -0.9582w + 0.9700$$

Thus, unsurprisingly, $i_F(0.5) = 0.4909$ for Table 2; exactly half its $i_F = 0.9818$ value. This weighted Finley index can also be written as a linear function of n_{11} by:

$$i_F(\mathrm{n}_{11}|w) = \frac{\mathrm{n}_{11}}{\mathrm{n}} + (1-w)\left(\frac{\mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1}}{\mathrm{n}}\right).$$

However, this relationship shows that the change in the index with respect to n_{11} is very small – only $1/\mathrm{n}$ since the sample size for each of Finley's data sets is quite large; corroborating the property that the impact of the cell frequencies on the index is negligible. This expression of the weighted Finley index also shows that it is under the influence more by the intercept than the slope, an intercept which is dependent only on the marginal frequencies and the choice of w. In fact, since $\mathrm{n}_{2\bullet} \gg \mathrm{n}_{\bullet 1}$ for Finley's data the intercept will be a little less than $1 - w$ so that:

$$i_F(w) \lesssim -w + 1$$

for all values of n_{11}. This again shows that the magnitude of the index will be influenced more by the choice of w than by any of the elements of the 2×2 contingency table. For Table 2, this relationship is:

$$i_F(\mathrm{n}_{11}|w) = 0.0011\mathrm{n}_{11} + 0.9582(1 - w) \approx -w + 1.$$

For example,

$$i_F(\mathrm{n}_{11}|0.6) = 0.0011\mathrm{n}_{11} + 0.5749$$

Which is depicted by the red dashed line in Figure 3. Therefore, by weighting the (1,1)th and (2,2)th cell frequencies of Finley's data like we have, this shows that $i_F(w)$ has no additional benefit when compared with i_F other than to reduce the scale of the index by about $-w + 1$. To further assess whether $i_F(w)$ is of greater utility than Finley's index, $i_F(w)$ is bounded by:

$$0 \leq \mathrm{L}_{F|w} = \frac{(1 - w)(\mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1})}{\mathrm{n}} < i_F(w) <$$

$$\frac{\min(\mathrm{n}_{1\bullet}, \mathrm{n}_{\bullet 1})}{\mathrm{n}} + \frac{(1 - \mathrm{w})(\mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1})}{\mathrm{n}} = \mathrm{U}_{F|w} < 1$$

Since $(\mathrm{n}_{2\bullet} - \mathrm{n}_{\bullet 1})/\mathrm{n} \approx 1$ and with $\mathrm{n} \geq \mathrm{n}_{1\bullet}$ and $\mathrm{n} \geq \mathrm{n}_{\bullet 1}$ for Finley's data then $\mathrm{L}_{F|w}$ and $\mathrm{U}_{F|w}$ will both be approximately $-w + 1$. This result can also be obtained from $i_F(\mathrm{n}_{11}|w)$ by also noting that $0 \lesssim \mathrm{n}_{11}/\mathrm{n}$ for Finley's data.

If the advice of Janson & Vegelius (1981) is followed so that n_{22} is ignored from the calculation of the bounds of $i_F(w)$ then $w = 1$ and:

$$\mathrm{L}_{F|1} = 0 < i_F(1) = \frac{\mathrm{n}_{11}}{\mathrm{n}} < \frac{\min(\mathrm{n}_{1\bullet}, \mathrm{n}_{\bullet 1})}{\mathrm{n}} = \mathrm{U}_{F|1}$$

which coincides with the Fréchet bounds of n_{11} when analysing Finley's data. However, if equal weights are given to n_{11} and n_{22} so that $w = 0.5$ then for Table 2:

$$\mathrm{L}_{F|0.5} = 0.4732 < i_F(0.5) < 0.4945 = \mathrm{U}_{F|0.5}$$

which, like $i_F \in [0.92,\ 0.97]$, is a very narrow interval with $i_F(0.5) = 0.4909$ lying very close to its upper bound. When $w = 0.6$ then $i_F(0.6)$ is bounded by

$$\mathrm{L}_{F|0.6} = 0.3833 < i_F(0.6) < 0.3987 = \mathrm{U}_{F|0.6}\,.$$

These results suggest that weighting n_{11} and n_{22} differently does not have any practical impact on the magnitude of $i_F(w)$ and the cell frequencies have very little bearing on it either. Therefore, irrespective of the choice of w, there seems to be little advantage in using $i_F(w)$ as an alternative to i_F.

7.4 A Weighted Finley Index (Version 2)

The second simple version of i_F that allows for the incorrect predictions that Finley made to be included is to define it so that:

$$j_F(w) = \frac{w(n_{11} + n_{22}) + (1 - w)(n_{12} + n_{21})}{n}.$$

for $w \in [0, 1]$. When $w = 1$ then $j_F(1) = i_F$ while $w = 0.5$ gives $j_F(0.5) = 0.5$. Otherwise, this index means that n_{12} and n_{21} are allowed to influence the magnitude of the index. This is a potential benefit since it means that less emphasis can be placed on n_{22} than when calculating i_F. However, this is at the cost of also placing less emphasis on n_{11}.

The index $j_F(w)$ can be expressed as a function of w, so that:

$$j_F(w) = w\left[1 - \frac{2(n_{1\bullet} + n_{\bullet 1} - 2n_{11})}{n}\right] + \frac{n_{1\bullet} + n_{\bullet 1} - 2n_{11}}{n}.$$

Since $(n_{1\bullet} + n_{\bullet 1} - 2n_{11})/n \approx 0$ for all four of Finley's data sets, an approximation of this function is:

$$j_F(w) \approx w$$

irrespective of the value of any of the cells of the contingency table . This version of the weighted Finley index can also be expressed as a function of n_{11} by

$$j_F(n_{11}|w) = -\frac{2(1 - 2w)}{n}n_{11} + \left[w + \frac{(1 - 2w)(n_{1\bullet} + n_{\bullet 1})}{n}\right].$$

Again, since n is large, then $-2(1 - 2w)/n \lesssim 0$ so that the slope of this function is approximately zero. If $w < 0.5$ then $-2(1 - 2w)/n \lesssim 0$ while $-2(1 - 2w)/n \gtrsim 0$ when $w > 0.5$. Since $(n_{1\bullet} + n_{\bullet 1})/n \lesssim 1$ for Finley's data, this function has an intercept of, approximately, w. Thus, $j_F(w) \lesssim w$ when $w > 0.5$ and $j_F(w) \gtrsim w$ when $w < 0.5$. For example, analysing Table 2 when $w = 0.3$ gives the function:

$$j_F(0.3) = -0.0009n_{11} + 0.3167 \gtrsim 0.3$$

so that $j_F(0.3) \in [0.3047, 0.3167]$, while

$$j_F(0.6) = 0.0004n_{11} + 0.5916 \lesssim 0.6$$

so that $j_F(0.6) \in [0.5916, 0.5972]$; a plot of n_{11} versus $j_F(0.6)$ is depicted by the green dashed line in Figure 4. In fact, for $w \in [0, 1]$, the range of values for the slope of $j_F(n_{11}|w)$ when analysing Table 2 is $[-0.00214, 0.00214]$ and is exactly zero at $w = 0.5$, while the range of its intercept values is $[0.0418, 0.9582]$ and whose values correspond, approximately, to the choice of $w \in [0, 1]$. Therefore, like $i_F(w)$, there seems to be little advantage in using $j_F(w)$ as an alternative to i_F.

7.5 Features of Gilbert's v

As we discussed in Section 4.2, Gilbert (1884, p. 166) was concerned that the prediction of a tornado occurring or not was "classed together as of equal difficulty". In proposing his index, v, Gilbert was also cognisant of the range of values it could take stating:

> "If [n_{11}, $n_{1\bullet}$ and $n_{\bullet 1}$] are numerically identical, it is evident that the ratio of verifications will be unity. If $[n_{11}] = 0$, the ratio of verification is also 0. Between these limits fall all practical cases"

Unlike Finley, Gilbert was thus aware of the bounds of his v even for the extreme cases when $n_{11} = n_{1\bullet} = n_{\bullet 1}$ and $n_{11} = 0$. A more general set of bounds for v can be obtained using the bounds of $n_{11} \in [L, U]$ and are:

$$L_v = \frac{\max(0, n_{\bullet 1} - n_{2\bullet})}{n_{1\bullet} + \min(n_{1\bullet}, n_{2\bullet})} < v < \frac{\min(n_{1\bullet}, n_{\bullet 1})}{\max(n_{1\bullet}, n_{\bullet 1})} = U_v \ .$$

However, since $\min(n_{1\bullet}, n_{2\bullet}) = n_{1\bullet}$ and $n_{\bullet 1} \ll n_{2\bullet}$ for Finley's data, the lower limit of v simplifies to $L_v = 0$ while $U_v = 1$ if and only if $n_{1\bullet} = n_{\bullet 1}$, otherwise $U_v < 1$. Therefore, for Table 2, v is bounded by $[0, 0.5600]$. Since $v = 0.3929$ for this data, it suggests that the ratio of verification is quite high, even keeping in mind that the maximum possible value it can take is 0.56 and not 1. For the March, May (8-hour) and May (10-hours) observations, $v \in [0, 0.3023]$, $v \in [0, 0.7143]$ and $v \in [0, 0.4545]$, respectively.

Since $n_{11} \in [0, 14]$ for Table 2, Gilbert's v can be expressed as a function of n_{11} so that:

$$v(n_{11}) = \frac{n_{11}}{39 - n_{11}}$$

which is depicted by dark blue dashed line in Figure 4. Under independence, Gilbert's v is:

$$v_{|I} = \frac{n_{1\bullet} n_{\bullet 1}}{n(n_{1\bullet} + n_{\bullet 1}) - n_{1\bullet} n_{\bullet 1}}$$

so that:

$$v_{|I} = \frac{25 \cdot 14}{934 \cdot (25 + 14) - 25 \cdot 14} = 0.0097$$

for Table 2. Comparing this small value with its observed value highlights that Gilbert's index appears to be more appropriate index on which to assess the verification of the occurrence of a tornado than Finley's index or its two weighted versions described in Sections 7.3 and 7.4. The shape of Gilbert's ratio of verification is certainly more aligned to the shape of Pearson's chi-squared statistic (Figure 3) that the Finley based indices.

7.6 Features of Gilbert's i_G

Since Gilbert amended his ratio of verification, v, by subtracting e from its numerator and denominator, its bounds are an emendation of $[\mathrm{L}_v, \mathrm{U}_v]$ and are:

$$\mathrm{L}_G = \frac{\max(0, \mathrm{n}_{\bullet 1} - \mathrm{n}_{2\bullet}) - e}{\mathrm{n}_{1\bullet} + \min(\mathrm{n}_{\bullet 1}, \mathrm{n}_{2\bullet}) - e} < i_G < \frac{\min(\mathrm{n}_{1\bullet}, \mathrm{n}_{\bullet 1}) - e}{\max(\mathrm{n}_{1\bullet}, \mathrm{n}_{\bullet 1}) - e} = \mathrm{U}_G\ .$$

Since $\mathrm{n}_{\bullet 1} \ll \mathrm{n}_{2\bullet}$ for Finley's data, these bounds simplify to:

$$0 < \mathrm{L}_G = -\frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}^2 - \mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 2}} < i_G < \min\left(\frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 2}}{\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 1}}, \frac{\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 2}}\right) = \mathrm{U}_G \leq 1$$

so that the minimum bound is always negative. Thus, $i_G \in [-0.0097,\ 0.5533]$ for Table 2 so that there is very little difference between this interval and the bounds of $v \in [0, 0.5600]$. Thus, subtracting e from the numerator and denominator does not greatly impact the bounds of i_G when compared with the bounds of v. We can also obtain the bounds of i_G for the March, May (8-hours) and May (10-hours) observations; they are $i_G \in [-0.0131,\ 0.2904]$, $i_G \in [-0.0106,\ 0.7091]$ and $i_G \in [-0.0129,\ 0.4443]$, respectively. Therefore, removing what he described as the "fortuitous coincidences" (e) from the numerator and denominator of v has not greatly impacted the magnitude of i_G for the March and April observations but it has affected the values that i_G can take for the two May data sets. Despite this, since independence between the variables results in $i_G = 0$, we posit that the observed value of $i_G = 0.3846$, being more similar to its upper bound than its value at independence, provides sufficient evidence to declare that Gilbert's index is a more suitable index for tornado prediction purposes than Finley's index.

Since Gilbert proposed a revision of his v index resulting in i_G, it is then of interest to compare the two indices across the range of $\mathrm{n}_{11} \in [\mathrm{L}, \mathrm{U}]$ values. While the lower bound of v, L_v, is always zero, $\mathrm{L}_G \lesssim 0$. Also, since $\mathrm{n}_{2\bullet} \approx \mathrm{n}_{\bullet 2}$ for all four of Finley's data sets then $\mathrm{U}_G \approx \mathrm{U}_v = 1$; we have commented in the previous paragraph that the two bounds are near identical for Table 2. The general shape of the two indices can be compared by observing that the first derivative of v and i_G with respect to n_{11} is:

$$\frac{\mathrm{d}}{\mathrm{dn}_{11}} v = \frac{\mathrm{n}_{12} + \mathrm{n}_{21}}{(\mathrm{n}_{1\bullet} + \mathrm{n}_{\bullet 1} - \mathrm{n}_{11})^2}$$

and

$$\frac{\mathrm{d}}{\mathrm{dn}_{11}} i_G = \frac{\mathrm{n}_{12} + \mathrm{n}_{21}}{(\mathrm{n}_{1\bullet} + \mathrm{n}_{\bullet 1} - \mathrm{n}_{11} - e)^2},$$

respectively. Since e is very small for Table 2 ($e = 0.3747$) – and also for the other three of Finley's data sets – then it has a negligible effect on the derivative of i_G. Thus, the behaviour of i_G and v are virtually identical for analysing Finley's data. This can also be seen by observing how close the two blue lines of Figure 4, representing the two indices, lie to each other. While there are practical benefits in removing e from the numerator and denominator of v doing so has virtually no impact on the value of the two indices for Table 2.

7.7 Features of Peirce's i_P

Suppose we now turn our attention to the features of Peirce's index and assess its suitability for assessing tornado predictions. Peirce's index can be alternatively expressed as:

$$i_P(\mathrm{n}_{11}) = \mathrm{n}_{11}\left(\frac{1}{\mathrm{n}_{\bullet 1}} + \frac{1}{\mathrm{n}_{\bullet 2}}\right) - \frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 2}}$$

so that, like Finley's index, $i_P$is linear function of n_{11}. For example, for Table 2

$$i_P(\mathrm{n}_{11}) = 0.0725\mathrm{n}_{11} - 0.02717$$

which is depicted dashed grey line in Figure 4. Peirce's index is bounded by:

$$\mathrm{L}_G = \max\left(0, \frac{(\mathrm{n}_{\bullet 1} - \mathrm{n}_{2\bullet})(\mathrm{n}_{\bullet 2} + \mathrm{n}_{1\bullet})}{\mathrm{n}_{1\bullet}\mathrm{n}_{2\bullet}}\right) - \frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 2}} < i_P < \min\left(\frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 1}}, 1 + \frac{\mathrm{n}_{\bullet 1} - \mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 2}}\right) = \mathrm{U}_G\,.$$

Since $\mathrm{n}_{\bullet 1} - \mathrm{n}_{2\bullet} < 0$ when studying Finley's data and $\mathrm{n}_{\bullet 1} - \mathrm{n}_{1\bullet}$ is very small relative to the very large $\mathrm{n}_{\bullet 2}$, then the bounds can be simplified giving:

$$\mathrm{L}_G = -\frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 2}} < i_P < \min\left(\frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 1}}, 1\right) \approx \mathrm{U}_G$$

so that the exact lower bound of i_P is always negative. A further approximation of these bounds can be made by noting that, when compared with the four sample sizes of Finley's data $\mathrm{n}_{\bullet 2} \gg \mathrm{n}_{1\bullet}$ and (for Table 2) $\mathrm{n}_{1\bullet} \approx \mathrm{n}_{\bullet 1}$. Thus, like Gilbert's index:

$$\mathrm{L}_P \approx 0 < i_P < 1 \approx \mathrm{U}_G.$$

For example, the bounds of i_P for Table 2 are $i_P \in [-0.0272, 0.9880] \approx [0, 1]$ as expected; see also Figure 4. For the March, May (10-hours) observations/predictions the bounds of i_P all approximately $[0, 1]$; $[-0.0567, 0.9604]$ and $[-0.0415, 0.9774]$ respectively. Although this approximation of the bounds is not always satisfied since $i_P \in [-0.0184, 0.7143]$ for the May (8-hours) data set.

Suppose we now assess Peirce's index under the assumption that the row and column variables of a 2×2 contingency table are independent. Then:

$$i_{P|I} = \frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}}\left(\frac{1}{\mathrm{n}_{\bullet 1}} + \frac{1}{\mathrm{n}_{\bullet 2}}\right) - \frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 2}}$$

which, after further simplification, is equal to zero.

7.8 Features of Doolittle's s_D

To investigate the features of Doolittle's s_D we can see that the parabolic relationship it has with n_{11} yields a function that has positive concavity. For Table 2 this relationship is

$$s_D(\mathrm{n}_{11}) = 0.0029\mathrm{n}_{11}^2.$$

and this is depicted by the dashed pink line in Figure 4. More generally, the first derivative of s_D with respect to n_{11} is:

$$\frac{\mathrm{d}}{\mathrm{dn}_{11}} s_D = \frac{2\mathrm{n}_{11}}{\mathrm{n}_{\bullet 1}\mathrm{n}_{1\bullet}}$$

so that its turning point exists at $\mathrm{n}_{11} = 0$ which is also the minimum value that n_{11} can take for the four data sets of Finley's data in Figure 1. The concavity of s_D is:

$$\frac{\mathrm{d}^2}{\mathrm{dn}_{11}^2} s_D = \frac{2}{\mathrm{n}_{\bullet 1}\mathrm{n}_{1\bullet}} > 0$$

so that the turning point of s_D coincides with its minimum value of zero at $\mathrm{n}_{11} = 0$. This second derivative also shows that the shape of this relationship is only dependent on $\mathrm{n}_{\bullet 1}$ and $\mathrm{n}_{1\bullet}$ so that the sample size and the large (2, 2)th value have no bearing on its shape. The quadratic relationship also suggests that there are two local maxima, although since the minimum exists at the lower bound the global maximum lies at the upper bound of the interval:

$$\mathrm{L}_s = 0 < s_D < \min\left(\frac{\mathrm{n}_{1\bullet}}{\mathrm{n}_{\bullet 1}}, \frac{\mathrm{n}_{\bullet 1}}{\mathrm{n}_{1\bullet}}\right) = \mathrm{U}_s$$

where the index will have a maximum of 1 only when $\mathrm{n}_{1\bullet} = \mathrm{n}_{\bullet 1}$, otherwise, $s_D < 1$. For Table 2, the range of values that s_D can take is:

$$\mathrm{L}_s = 0 < s_D < 0.5600 = \mathrm{U}_s\,.$$

If predicting the number of tornadoes happens completely by chance, then Doolittle's index simplifies to:

$$s_{D|I} = \frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}^2}$$

and is the expected value of the proportion of successfully predicted tornadoes, $p_{11} = n_{11}/n$, when the rows and columns of the 2 × 2 table are completely independent. Therefore, under independence Doolittle's index for Table 2 is:

$$s_{D|I} = \frac{25 \cdot 14}{934^2} = 0.0004$$

which lies very close to the smallest value of n_{11} assuming the marginal frequencies of Table 2 are known.

7.9 Features of Doolittle's i_D

Like s_D, Doolittle's i_D is a quadratic function of n_{11} with positive concavity. For Table 2, this function is:

$$i_D(n_{11}) = 0.00298(n_{11} - 0.3747)^2$$

which, upon expansion, is (for all practical purposes) identical to $s_D(n_{11})$. This (near) equivalency can be seen by observing that the solid yellow line depicting $i_D(n_{11})$ in Figure 4 lies (with the slightest of deviations) on top of the line depicting $s_D(n_{11})$. In fact, since $X^2(n_{11}) \propto i_D(n_{11})$, the yellow solid line in Figure 4 is identical in shape to the curve depicted in Figure 3.

In more general terms, the first and second derivative of i_D with respect to n_{11} is:

$$\frac{d}{dn_{11}} i_D = \frac{2n^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}\left(n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{n}\right)$$

and

$$\frac{d^2}{dn_{11}^2} i_D = \frac{2n^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}}$$

respectively. Therefore, the minimum value that i_D can take is when there is independence between the row and column variables of the 2 × 2 table; that is, when $n_{11} = e = 0.3750$ for Table 2 (this feature is shared with s_D). Thus, there are two local maxima which lie at the bounds:

$$L_D = \frac{\{\min(n_{1\bullet}n_{\bullet 1}, n_{2\bullet}n_{\bullet 2})\}^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}} < i_D < \frac{\{\min(n_{1\bullet}n_{\bullet 2}, n_{2\bullet}n_{\bullet 1})\}^2}{n_{1\bullet}n_{\bullet 1}n_{2\bullet}n_{\bullet 2}} = U_D .$$

These bounds can be simplified further to:

$$0 \lesssim L_D = \min\left(\frac{n_{1\bullet}n_{\bullet 1}}{n_{2\bullet}n_{\bullet 2}}, \frac{n_{2\bullet}n_{\bullet 2}}{n_{1\bullet}n_{\bullet 1}}\right) < i_D < \min\left(\frac{n_{1\bullet}n_{\bullet 2}}{n_{2\bullet}n_{\bullet 1}}, \frac{n_{2\bullet}n_{\bullet 1}}{n_{1\bullet}n_{\bullet 2}}\right) = U_D < 1.$$

Since $n_{1\bullet} \ll n_{2\bullet}$ and $n_{1\bullet} \ll n_{2\bullet}$ for Finley's data then:

$$0 \lesssim \mathrm{L}_D = \frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 2}} < i_D < \min\left(\frac{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 2}}{\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 1}}, \frac{\mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 1}}{\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 2}}\right) = \mathrm{U}_D < 1\,.$$

so that the global maximum of i_D lies at the upper bound of this interval. Note that the upper bound of i_D is identical to the upper bound of i_G for all 2×2 contingency tables and not just for Finley's data. For Table 2:

$$\mathrm{L}_D = 0.0004 < i_D < 0.5533 = \mathrm{U}_D$$

where the (minimum) turning point coincides with independence that lies near close to the lower bound.

Suppose we now compare the features of i_D and s_D. We first note that the bounds of both indices are approximately the same; $\mathrm{L}_s \lesssim \mathrm{L}_D$ since $\mathrm{n}_{1\bullet}\mathrm{n}_{\bullet 1} \ll \mathrm{n}_{2\bullet}\mathrm{n}_{\bullet 2}$ while $\mathrm{U}_s \approx \mathrm{U}_D$ when $\mathrm{n}_{2\bullet} \approx \mathrm{n}_{\bullet 2}$ which is certainly the case for Finley's data. A comparison of the concavity of i_D and s_D can be made by observing that:

$$\frac{\mathrm{d}^2}{\mathrm{dn}_{11}^2} i_D = \left(\frac{\mathrm{d}^2}{\mathrm{dn}_{11}^2} s_D\right)\left(\frac{\mathrm{n}}{\mathrm{n}_{2\bullet}} \cdot \frac{\mathrm{n}}{\mathrm{n}_{\bullet 2}}\right).$$

Since $\mathrm{n}_{2\bullet} \lesssim \mathrm{n}$ and $\mathrm{n}_{\bullet 2} \lesssim \mathrm{n}$ then:

$$\frac{\mathrm{d}^2}{\mathrm{dn}_{11}^2} s_D \lesssim \frac{\mathrm{d}^2}{\mathrm{dn}_{11}^2} i_D$$

and are equivalent when all of the sample is allocated into the second row and second column categories (which was not observed in Finley's predictions/observations and is not assumed to be a legitimate allocation of marginal frequencies). For Table 2, $\mathrm{d}^2 i_D/\mathrm{dn}_{11}^2 = 0.00596$ and $\mathrm{d}^2 s_D/\mathrm{dn}_{11}^2 = 0.00571$ and so, since the bounds, and shape of Doolittle's two indices are near equivalent, their behaviour across the interval of n_{11} values is approximately the same. This suggests that, while Doolittle felt compelled to obtain his "vitiated probability", there was no practical reason for doing so. Figure 4 shows that the behaviour of s_D and i_D across the interval $\mathrm{n}_{11} \in [0, 14]$ are near equivalent.

7.10 Which Index?

An obvious and reasonable question that one may ask at this point is *which index should be most preferred*?

If Pearson's mean square contingency is the benchmark on which any evaluation is based, then it should now be clear that Doolittle's i_D is the index of choice for analysing ANY 2×2 contingency tables since $i_D = \phi^2$. However, if we were to restrict ourselves to the analysis of Finley's data then Doolittle's s_D works equally well since $e \approx 0$. Another advantage of using Doolittle's i_D is that it is equivalent to Cohen's kappa (Cohen, 1960); see Armistead (2016) for more of a discussion of this equivalency.

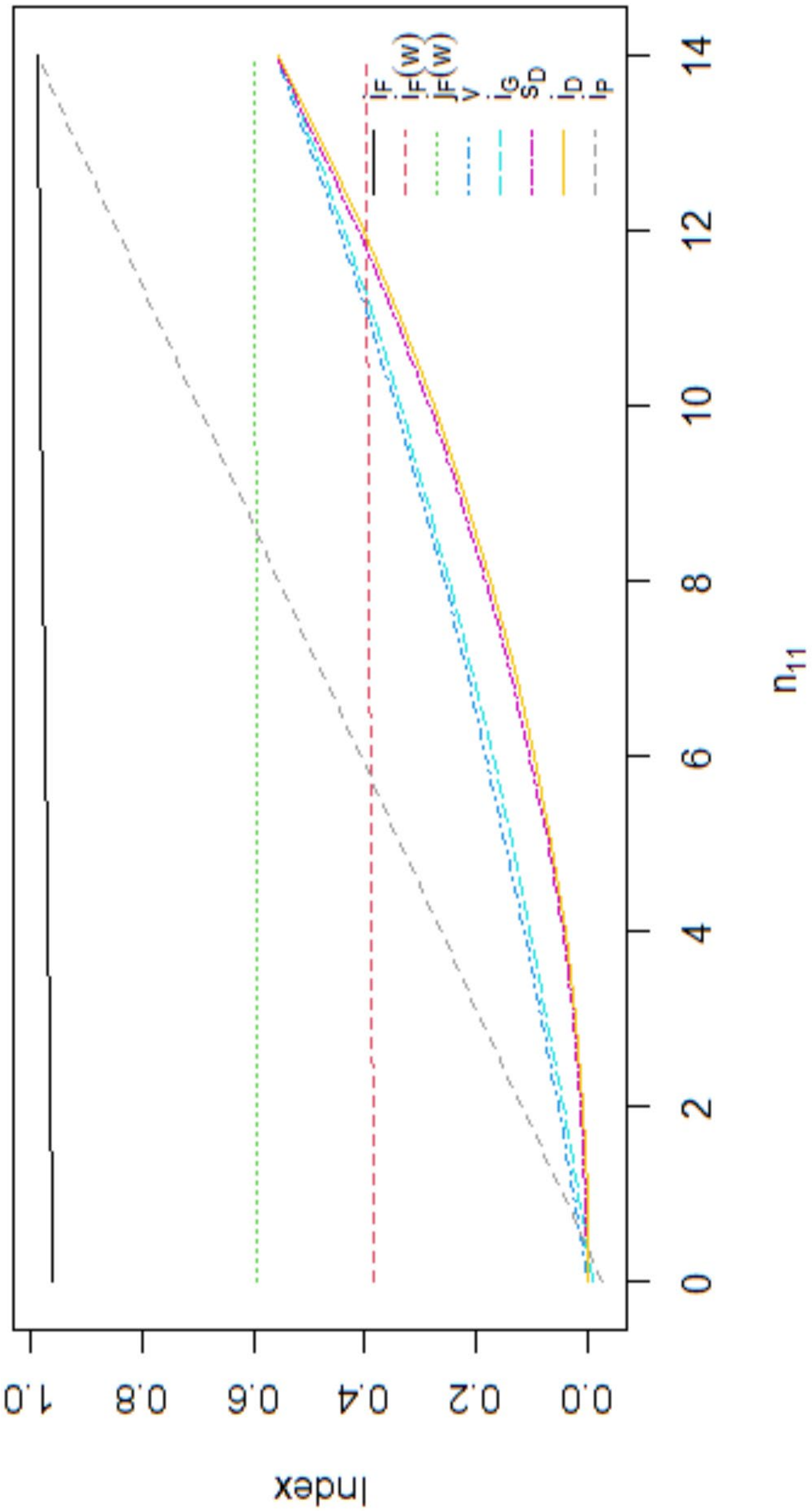

FIGURE 4. Plot of the n_{11} versus i_F, $i_F(0.6)$, $j_F(0.6)$, v, i_G, s_G, i_D and i_P for Finley's April predictions and observations (Table 2)

Table 3. Value of each observed index, its value under independence and the interval of possible values it takes for the four data sets and their aggregation in Figure 1

Index/Month	Index	Independence	Bounds
i_F			
March	0.9429	0.9292	[0.9274, 0.9611]
April	0.9818	0.9590	[0.9582, 0.9882]
May (8 hours)	0.9857	0.9579	[0.9570, 0.9928]
May (10 hours)	0.9519	0.9422	[0.9407, 0.9778]
Aggregate	0.9661	0.9474	[0.9461, 0.9825]
$i_F(0.6)$			
March	0.3787	0.3719	[0.3709, 0.3878]
April	0.3951	0.3837	[0.3833, 0.3983]
May (8 hours)	0.3971	0.3832	[0.3828, 0.4007]
May (10 hours)	0.3819	0.3771	[0.3763, 0.3948]
Aggregate	0.3884	0.3791	[0.3785, 0.3966]
$j_F(0.6)$			
March	0.5886	0.5858	[0.5855, 0.5922]
April	0.5964	0.5918	[0.5916, 0.5976]
May (8 hours)	0.5971	0.5916	[0.5914, 0.5986]
May (10 hours)	0.5904	0.5884	[0.5881, 0.5956]
Aggregate	0.5932	0.5895	[0.5892, 0.5965]
v			
March	0.1200	0.0131	[0, 0.3023]
April	0.3929	0.0097	[0, 0.5600]
May (8 hours)	0.5000	0.0106	[0, 0.7143]
May (10 hours)	0.1034	0.0129	[0, 0.4545]
Aggregate	0.2276	0.0122	[0, 0.5100]
i_G			
March	0.1071	0	[-0.0131, 0.2904]
April	0.3846	0	[-0.0097, 0.5533]
May (8 hours)	0.4920	0	[-0.0106, 0.7091]
May (10 hours)	0.0907	0	[-0.0129, 0.4443]
Aggregate	0.2160	0	[-0.0122, 0.5009]
i_P			
March	0.4127	0	[-0.0567, 0.9604]
April	0.7705	0	[-0.0272, 0.9880]
May (8 hours)	0.5678	0	[-0.0184, 0.7143]
May (10 hours)	0.2642	0	[-0.0415, 0.9774]
Aggregate	0.5229	0	[-0.0363, 0.9822]
s_D			
March	0.0644	0.0009	[0, 0.3023]
April	0.3457	0.0004	[0, 0.5600]
May (8 hours)	0.4571	0.0004	[0, 0.7143]
May (10 hours)	0.0409	0.0008	[0, 0.4545]
Aggregate	0.1537	0.0006	[0, 0.5100]
i_D			
March	0.0536	0	[0.0001, 0.2904]
April	0.3365	0	[0.0004, 0.5533]
May (8 hours)	0.4480	0	[0.0005, 0.7091]
May (10 hours)	0.0325	0	[0.0008, 0.4443]
Aggregate	0.1420	0	[0, 0.5009]

Gilbert's v and i_G work well for Finley's data with the latter of these two being preferable in general since it incorporates any deviation n_{11} has from what is expected under independence (e). Recall that, like Doolittle's two indices, both of Gilbert's indices behave in a very similar way since $e \approx 0$ for Table 2.

Peirce's index performs admirably for gaining a broad understanding of the association and has the features that Yule espoused for assessing association. However, given the four options just discussed, there are better indices available. Although Baker and Kramer (2007) do point out that the maximum of i_P coincides with the optimal point on a receiver operating characteristic curve and so has utility in that context.

Despite the energy that followed the publication of Finley's data and his index, and the excellent emendations that were made, it should be clear that his index is to be avoided. While this paper has made a point of ensuring that there may be options other than those presented in Sections 7.3 and 7.4 for better weighting the cells of a 2×2 contingency table, the weighted versions, $i_F(w)$ and $j_F(w)$, are also to be avoided.

8 Curtis (1887) and His Concerns

Before ending our discussion, we provide some brief context on some concerns raised by George E. Curtis (1887), a member of the US Signal Service before working at Washburn College, Kansas, USA (Frazier & Heckler, 1972, p. 8). While we showed in Section 7.6 that subtracting e from the numerator and denominator of v, as Gilbert (1884) did (resulting in i_G), is not of practical use when analysing Finley's data, Curtis (1887) agreed it was generally wise to do so saying:

> "To obtain a measure of the skill in prediction, the portion of the total success that is due to probable accidental coincidences must be eliminated, for only by doing so can the success of predictions of different phenomena be rendered comparable."

However, Curtis (1887) was critical of the assumption Finley and Gilbert implicitly made that tornadoes would occur uniformly across all districts adding:

> "To obtain e when the predictions are made for more than one district, the formula requires an extension due to the fact that the occurrences are not distributed uniformly over the several districts, but have a relative frequency now fairly well determined . . . The relative frequency being known, the predictions might be given the same distribution by a sunspot prediction or mere guess. The value of e, therefore, must be computed for each district separately".

Curtis (1887, p. 70) then suggests that the overall expected number of correct tornado predictions across all d, say, districts should be calculated by:

$$e = \frac{n_{1\bullet 1} n_{\bullet 11}}{n} + \frac{n_{1\bullet 2} n_{\bullet 12}}{n} + \cdots + \frac{n_{1\bullet d} n_{\bullet 1d}}{n} = \sum_{k=1}^{d} \frac{n_{1\bullet k} n_{\bullet 1k}}{n}$$

where $n_{1\bullet k}$ is the number of tornadoes predicted to have occurred in the kth district and $n_{\bullet 1k}$ is the number of tornadoes that occur in that district. Note here that Curtis (1887) calculates the expected number of correctly predicted tornadoes for each district separately BUT does so by dividing by sample size of all districts rather than the sample size of each district. If he had done so, the expectation would be defined by:

$$e = \sum_{k=1}^{d} \frac{n_{1\bullet k} n_{\bullet 1k}}{n_k}$$

where n_k is the sample size of the kth district and this version of the global expected value can be seen in the Mantel-Haenzel statistic (Mantel & Haenszel, 1959, p. 736).

Not only did Curtis (1887) question the appropriateness of Gilbert's calculation of *e* but he also thought that:

> ". . . the predictions of tornadoes for fixed districts is uneconomical" (p. 71)

So, Curtis proposed that there should be an aggregation/pooling of some districts since:

> ". . . not only are the districts too small, but the whole system of fixed districts is inappropriate for making advantageous predictions" (p. 72).

This was despite Finley's (1884b, p. 86) recommendation that:

> "Tornadoes being remarkably local disturbances, the area for which predictions are made on any one day should be as limited as possible".

Curtis (1887) then proposes to aggregate some of the districts and suggested that such aggregations should be of a certain shape and size. He recommended that the shape of any district designed for tornado prediction and verification should either be rectangular where its length is 1.5 times its breadth or be elliptical where the semi-major and semi-minor axis lengths are of the same dimension. Regarding the size of the districts, they should be of dimension 400x600 miles, or about 645x966 kilometres.

Some of the statistical implications of spatial aggregation were raised much later by Gehlke & Biehl (1934) while Holt, Steel, Tranmer & Wrigley (1996) provide an excellent discussion of the history and some of the key issues. Entwined in spatial aggregation is the aggregation of data. The problems encountered with this type of aggregation, especially when it concerns stratified 2×2 contingency tables, is an ongoing issue in *ecological inference* (EI). EI is an area of contingency table analysis that is concerned with estimating the cell-counts (or some function of them) in stratified 2×2 contingency tables when all that is available for analysis is the row and column margins of each stratum, or in Finley's case, districts. Generally, the methods developed for performing EI have relied extensively on Bayesian techniques and one may refer to Goodman (1953), King (1997, 2004), Chambers & Steel (2001), Kousser (2001), King, Rosen & Tanner (2004), Wakefield (2004), Lau, Moore & Kellerman (2007), Hudson, Moore, Beh & Steel (2010), Imai, Lu & Strauss (2011) and Knudson, Schoenbach & Becker (2021), and as excellent examples of the history, issues, solutions and computer packages available for performing EI. If one does not wish to make assumptions about the underlying nature of the data, or features of these techniques, the association structure of the variables of a 2×2 contingency table can be assessed using the aggregate association index; see, for example, Beh (2010), Beh, Tran & Hudson (2013, 2024) and Lombardo & Beh (2016).

9 Discussion

9.1 A Recap on the Contributions

This paper has given some historical perspective to, and an evaluation of, the contributions of Finley and three of his successors to the analysis of contingency table. The timing is important since they occurred a decade or two prior to the work of Galton, Pearson and Yule – justifiably important figures in the origins of categorical data analysis and, more generally, statistical practice. Summing up much of the discussion made in the earlier sections of this paper, the statistical contributions of Finley, Gilbert, Peirce and Doolittle include:

- Their analysis of the 2×2 contingency table, and the derivation of the indices in (largely) general terms but with a practical flavour on the data in Figure 1,
- The awareness of Peirce (1884) that the indices described could be generalised to tables of size bigger than 2×2, even though such generalisations were not presented,
- Gilbert's (1884) quantification and rationalisation of e, the expected value of the (1, 1)th cell of the contingency table under the assumption of independence between the row and column variables,
- Gilbert's (1884) conception, interpretation, and usage of a difference measure Pearson (1904) would refer to as the "contingency" of a cell, and
- Doolittle's (1885) *degree of logical connection,* i_{D}, that is equivalent to Pearson's (1904) mean square contingency.

The impact made by our quartet (and others not included in this discussion) was especially profound with Dowsell, Weiss & Johns (1993) pointing out that their influence was so pervasive during the period spanning the 1890's to the early 20$^{\text{th}}$ century that it was regarded as the "dark age" of tornado prediction. This is because the tools that were becoming available to forecast tornados, or to verify how well these tools worked, were felt to be more harmful than helpful to the general public (Grice *et al.* 1999, p. 1345). Bradford (1999, p. 489) points out that in 1905 the US Weather Bureau Stations Regulations contained the statement

"Forecasts of tornadoes are prohibited" and this ban would be in place until 1938. Bradford (1999, p. 490) also added that

> "In spite of 4151 tornado deaths from 1920 to 1939, including 794 in 1925 alone, the Weather Bureau did nothing to try to reduce the loss of life from these natural disasters . . . the state of tornado forecasting and warnings was nonexistent in 1940 as it had been in 1870."

9.2 On the Lack of Attention Received from those in the UK

If these contributions to the meteorological literature had such importance in the evolving statistical literature an obvious question to ask is *"why aren't these early pioneers included more than they are in discussions on the early history of contingency table analysis?"*. There are certainly exceptions including Goodman and Kruskal (1959, Section 3.1), Rovine & Anderson (2004), Baker & Kraner (2007) and Armistead (2016) who discussed some aspects of the work of Finley, Gilbert, Peirce and Doolittle. However, the works described in this paper by our quartet are worthy of more consideration than they are given. There is no doubt a multitude of intermingled answers to this question, but a brief account is given of two rather broad possible reasons, especially given the state global scientific research at the time:

- *Use of Vernacular*: Throughout our discussion we have noted that much of the language used by Finley, Gilbert, Peirce and Doolittle to describe random events was dominated by terms such as "chance", "coincidence" and "fortuitous". Their use, like the work of Galton, Pearson and Yule, shows that Finley, Gilbert, Peirce and Doolittle were not concerned with causal links between variables but were very much interested in what we now understand to be "association". This may be simply because of the nature of the variables they studied – *Predicted* and *Occurrence* – do not permit a natural causal link to be investigated. Although this appears contrary to Pearson (1930, p. 1) who wrote:

  > "Up to 1889 men of science had thought only in terms of causation, in future they were to admit another working category, that of correlation . . .".

  when writing about Galton's development of correlation; see also Aldrich (1995, p. 365) for a detailed discussion on this issue. However, as we describe next, it is very possible that Pearson was speaking only on behalf of the "men of science" in the UK and not of those in the US or elsewhere. Very few statisticians and other allied practitioners of statistics today use the term's "chance" or "coincidence" in the literature (although they are often used when teaching introductory undergraduate service statistics courses), instead using "correlation" and "association". Despite this, the very essence of these (and other similar) terms aligns perfectly with how statisticians over the past 120 years have described the meaning of "correlation" and "association". It must be kept in mind that the difference in vernacular is not in how Finley *et al.* used their terms but in the timing of when "correlation" and "association" were first used. As it was noted in Section 6.2, "association" was first used by Yule (1900) while "correlation" was used by Galton (1888) and these terms have been used extensively throughout the statistical literature ever since. So, it very much seems that the statistical spirit of the work undertaken during the "Finley affair" was still very much alive even if they didn't use the same terms developed and adopted at the turn of the 20th century.

- *The Continental or Discipline Divide?*: It may be tempting to argue that the lack of awareness of the "Finley affair" by those in UK and Europe may be due to the distance that lies between the two continents so that there was a general lack of awareness of the work being done. Was it, instead, due to some bias that exists between them, or was there some other reason that meant that the likes of Galton and Pearson did not refer to the work of Finley and his successors? The first two sentences of Kelves, Sturchio & Carroll (1980, p. 26) may help to answer this question. They painted a vivid picture of the "continental biases" that existed between scientists in Europe and the in US at the turn of the 20th century (I am including the UK to be within the confines of Europe here). They said:

  > "For many years American science circa 1880 was understood to have been a primitive enterprise, a colonial outpost of European research, an intellectual backwater. The research of the time was written off as merely applied work and, hence, by some mysterious logic, as insignificant."

  Such a sentiment was shared around the same time as the "Finley affair" when US physicist Rowland (1883, p. 242) colourfully said of the state of science in the US:

  > "I go out to gather grain ripe to the harvest, and I find only tares. Here and there a noble head of grain rises above the weeds; but so few are they, that I find the majority of my countrymen know them not, but think that they have a waving harvest, while it is only one of weeds after all."

  Therefore, it seems highly likely that the work undertaken during the "Finley affair" was not viewed as being significant or original (or at the very least, being merely practical) in the eyes of scientists in Europe and the UK at the time. Therefore, it may even be plausible to suggest that Galton and Pearson were not aware of the work undertaken by Finley *et al.*. This may be suggested at since the work of those involved in the "Finley affair" appeared in US-centric publications (for example, *The American Meteorological Journal*, *Science* and *Bulletin of the Philosophical Society of Washington*) while the works of Galton and his successors appeared in UK-centric publications (such as *Philosophical Transactions of the Royal Society of London*, *Philosophical Magazine*, *Biometrika*, *Drapers' Company Research Memoirs* and *Journal of the Royal Statistical Society*). However, it is clear that Galton was familiar with at least some of what was published in *Science* (being a US-centric journal) having published a letter there (Galton, 1880). One also does not have to go too far into his *Finger Prints* book to see the evidence. For example, Galton (1892, p. 26) states "A correspondent of the American Journal *Science*, viii 166, . . ." in reference to Hough (1886) who discussed the use "thumb and finger markings" in China for identification purposes, and that Chinese pots were made with patterns resembling fingerprint features. Pearson's awareness of the activities in the US was also apparent. Bellhouse (2009) provides a very interesting account of Pearson's influence in the US and the contacts he maintained there, although these span the early part of the 20th century through to his death in 1936.

It therefore seems unlikely that Galton and Pearson were aware of the events underlying the Finley affair and that the people they were in contact with were biologically and not meteorologically focused. Irrespective of whether Galton and Pearson were aware of the Finley affair, it must be agreed that, while Finley *et al.* were insightful in their arguments and derivations, they focused primarily on resolving a specific practical problem and their work (perhaps with the exception of Doolittle) lacked the depth and rigour when compared with the contributions of Galton, Pearson and their successors. It should therefore be of no surprise that it is this depth and rigour that have endeared Galton, Pearson *et al.* to the statistical and allied literature.

9.3 Some Final Thoughts

The legacy of Finley, his data and his index are well documented and understood in the meteorological literature. The lack of attention given to Finley, Gilbert, Peirce and Doolittle is certainly not because their contributions are not without merit (they certainly are worthy of merit) but because of the impressive, thorough and rigorous manner in which Pearson, Yule and their successors went about developing the foundations of categorical data analysis 20 years later. It is to be hoped that this paper better places the contributions of Finley, Gilbert, Peirce and Doolittle in further discussions on the evolution of contingency table analysis and that any conversation on this topic includes something of their work.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed). Wiley.

Agresti, A. (2010). *Analysis of Ordinal Categorical* Data (2nd ed). Wiley.

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed). Wiley.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, **10**(4), 364 – 376.

Andersen, E. B. (1994). *The Statistical Analysis of Categorical Data* (3rd ed). Springer-Verlag.

Andersen, E. B. (1997). *Introduction to the Statistical Analysis of Categorical Data*. Springer.

Armatte, M. (2008). Histoire et Préhistoire de l'Analyse des données par J.P. Benzécri: un cas de généalogie rétrospective. *Electronic Journal for History of Probability and Statistics*, **4**(2), 1 – 24.

Armistead, T. W. (2016). Misunderstood and unattributed: Revisiting M. H. Doolittle's measures of association, with a note on Bayes theorem. *The American Statistician*, **70**(1), 63 – 73.

Azen, R. & Walker, C. M. (2021). *Categorical Data Analysis for the Behavioral and Social Sciences* (2nd ed). Routledge.

Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review*, **72**(2), 257 – 284.

Beh, E. J. (2010), The aggregate association index, *Computational Statistics & Data Analysis*, **54**, 1570 – 1580

Beh. E. J., Tran, D. & Hudson, I. L. (2013), A reformulation of the aggregate association index using the odds ratio, *Computational Statistics & Data Analysis*, **68**, 52 – 65.

Beh, E. J., Tran, D. & Hudson, I. L. (2024), A generalization of the aggregate association index (AAI): Incorporating a linear transformation of the cells of a 2×2 table, *Metrika* (in press)

Beh, E. J. & Lombardo, R. (2012). A genealogy of correspondence analysis. *Australian & New Zealand Journal of Statistics*, **54**(2), 137 – 168.

Beh, E. J. & Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. Wiley.

Beh, E. J. & Lombardo, R. (2019). A genealogy of correspondence analysis: Part 2 – the variants. *Electronic Journal of Applied Statistical Analysis*, **12**(2), 552 – 603.

Beh, E. J. & Lombardo, R. (2021). *An Introduction to Correspondence Analysis*. Wiley.

Bellhouse, D. R. (2009). Karl Pearson's influence in the United States. *International Statistical Review*, **77**(1), 51 – 63.

Bilder, C. R. & Loughlin, T. M. (2015). *Analysis of Categorical Data with R*. CRC Press

Blasius, J. & Greenacre, M. (eds) (1998). *Visualization of Categorical Data*. Academic Press.

Bradford, M. (1999). Historical roots of modern tornado forecasts and warnings. *Weather and Forecasting*, **14**(4), 484 – 491.

Chambers, R. L. & Steel, D. G. (2001). Simple methods for ecological inference in 2×2 tables. *Journal of the Royal Statistical Society, Series A*, **164**(1), 175 – 192.

Clogg, C. C. & Shihadeh, E. S. (1994). *Statistical Models for Ordinal Variables*. Sage Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37 – 46.

Congdon, P. (2005). *Bayesian Models for Categorical Data*. Wiley.

Cuadras, C. M. & Greenacre, M. (2022). A short history of statistical association: from correlation to correspondence analysis to copulas. *Journal of Multivariate Analysis*, **188**, 104901.

Curtis, G. E. (1887). Tornado predictions and their verification. *The American Meteorological Journal*, **4**, 68 – 74.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician*, **49**(2), 121 – 133.

David, H. A. (1998). First (?) occurrence of common terms in probability and statistics – a second list, with corrections. *The American Statistician*, **52**(1), 36 – 40.

De Leeuw, J. (1983). On the prehistory of correspondence analysis. *Statistica Neerlandica*, **37**(4), 161 – 164.

Diamond, M. & Stone, M. (1981), Nightingale on Quetelet, *Journal of the Royal Statistical Society, Series A (General)*, **144**(1), 66 – 79.

Doolittle, M. H. (1878). Method employed in the solution of normal equations and in the adjustment of a triangularization. In *Report of the Superintendent of the Coast and Geodetic Survey Showing the Progress of the Work During the Fiscal Year Ending with June, 1878*, pp. 115 – 120, Government Printing Office, Washington DC (Published 1881).

Doolittle, M. H. (1885). The verification of predictions. *Bulletin of the Philosophical Society of Washington*, **7**, 122– 127.

Doolittle, M. H. (1888). Association ratios. *Bulletin of the Philosophical Society of Washington*, **10**, 83 – 87 & 94 – 96.

Doswell III, C. A., Weiss, S. J. & Johns, R. H. (1993). Tornado forecasting: a review, In *The Tornado: Its Structure, Dynamics, Prediction, and Hazards* (eds Church, C., Burgess, D., Doswell, C. & Davies-Jones, R), pp. 557 – 571, American Geophysical Union.

Dwyer, P. S. (1941). The Doolittle technique. *The Annals of Mathematical Statistics*, **12**(4), 449 – 458.

Eden, F. (1797). *The State of the Poor, or, an History of the Labouring Classes in England from the Conquest to the Present Period* (Volume 2). J. Davis, London.

Eden, F. (1802). *Eight Letters on the Peace; and on the Commerce and Manufactures of Great Britain and Ireland*. Wilks and Taylor, London.

Everitt, B. S. (1992). T*he Analysis of Contingency Tables* (2nd ed). Chapman and Hall.

Fagerland, M. W., Lydersen, S. & Laake, P. (2017). *Statistical Analysis of Contingency Tables*. CRC Press.

Fienberg, S. E. (2007). *The Analysis of Cross-Classified Categorical Data* (2nd ed), Springer.

Finley, J. P. (1884a). Intelligence from American scientific stations. *Science*, **3**, 766 – 768.

Finley, J. P. (1884b). Tornado predictions. *The American Meteorological Journal*, **1**, 85 – 88.

Frazier, A. H. & Heckler, W. (1972). *Embudo, New Mexico, Birthplace of Systemic Stream Gaging*. Geological Survey Professional Paper 778.

Fréchet, M. (1951) Sur les tableaux de corrélation dont les marges sont données. *Review de l'Institut International de Statistique*, **28**, 10 – 32.

Friendly, M. (2000). *Visualizing Categorical Data*. SAS Institute.

Friendly, M. (2002). A brief history of the mosaic display. *Journal of Computational and Graphical Statistics*, **11**(1), 89 – 107.

Friendly, M. (2006). A brief history of data visualization. In *Handbook of Computational Statistics* (Chen, C., Härdle, W. & Unwin, A. eds), pp. 15 – 56, Springer.

Friendly, M. & Meyer, D. (2018). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. CRC Press.

G. (1884). Tornado predictions. *Science*, **4**(80), 126 – 127.

Galton, F. (1880). Mr Francis Galton's proposed 'family registers' (letter to the editor). *Science*, **4**, page 3.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, **45**, 135 – 145.

Galton, F. (1892). *Finger Prints*. MacMillan and Co., London.

Galway, J. G. (1985). J. P. Finley: the first severe storms forecaster. *Bulletin of the American Meteorological Society*, **66**(11), 1389 – 1395 & **66**(12), 1506 – 1510.

Galway, J. G. (1992). Early severe thunderstorm forecasting and research by the United States Weather Bureau, Weather and Forecasting, *Weather and Forecasting*, ***7***(4), 564 – 587.

Gehlke, C. E. & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, **29**(Supplement), 169 – 170.

Gilbert, G. K. (1884). Finley's tornado predictions. *The American Meteorological Journal*, **4**, 166 – 172.

Gillham, N. W. (2001). Sir Francis Galton and the birth of eugenics. *Annual Review of Genetics*, **35**, 83 – 101.

Gokhale, D. V. & Kullback, S. (1978). *The Information in Contingency Tables*. Marcel Dekker.

Goodman, L. A. (1953). Ecological regressions and the behavior of individuals. *American Sociological Review*, **18**(6), 663 – 664.

Goodman (1978). *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis*, Abt Books.

Goodman (1984). *The Analysis of Cross-Classified Data Having Ordered Categories*. Harvard University Press.

Goodman, L. A. & Kruskal, W. H. (1959). Measures of association for cross classifications II: Further discussions and references. *Journal of the American Statistical Association*, **54**, 123 – 163.

Goodman, L. A. & Kruskal, W. H. (1979). *Measures of Association for Cross Classifications*. Springer-Verlag.

Grcar, J. F. (2011). Mathematicians of Gaussian elimination. *Notices of the American Mathematical Society*, **58**(6), 782 – 792.

Graunt, J. (1662). *Natural and Political Observations, Mentioned in the following Index, and made upon the Bills of Mortality*. Roycroft, London.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Greenacre, M. (2017). *Correspondence Analysis in Practice* (3rd ed). CRC Press, Boca Raton, FL.

Grice, G. K., Trapp, R. J., Corfidi, S. F., Davies-Jones, R., Buonanno, C. C., Craven, J. P., Droegemeier, K. K., Duchon, C., Houghton, J. V., Prentice, R. A., Romine, G., Schlachter, K. & Wagner, K. K. (1999). The golden anniversary celebration of the first tornado forecast, *Bulletin of the American Meteorological Society*, **80**(7), 1341 – 1348.

Haberman, S. J. (1978). *Analysis of Qualitative Data, Volume 1: Introductory Topics*. Academic Press.

Haberman, S. J. (1979). *Analysis of Qualitative Data, Volume 2: New Developments*. Academic Press.

Haldane, J. B. S. (1957). Karl Pearson, 1857 – 1957. *Biometrika*, **44**(3/4), 303 – 313.

Hankins, F. H. (1908), *Adophe Quetelet as Statistician*, PhD Thesis, Faculty of Political Science, Columbia University (Reprinted in 1968 by AMS Press, New York).

Hill, M. O. (1973). Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology*, **61**(1), 237 – 249.

Hill, M. O. (1974). Correspondence analysis: A neglected multivariate method. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **23**(3), 340 – 354.

Hogan, R. J., Ferro, C. A. T., Jolliffe, I. T. & Stephenson, D. B. (2009). Equitability revisted: Why the 'equitable threat score' is not equitable. *Weather and Forecasting*, **25**(2), 710 – 726.

Holt, D., Steel, D. G., Tranmer, M. & Wrigley, N. (1996). Aggregation and ecological effects in geographically based data. *Geographical Analysis*, **28**(3), 244 – 261.

Hough, W. (1886). Thumb marks. *Science*, **8**, 166 – 168.

Hudson, I. L., Moore, L., Beh, E. J. & Steel, D. G. (2010). Ecological inference techniques: An empirical evaluation using data describing gender and voter turnout and New Zealand elections, 1893 – 1919. *Journal of the Royal Statistical Society, Series, A*, **173**(1), 185 – 213.

Imai, K., Lu, Y. & Strauss, A. (2011). eco: R package for ecological inference in 2 × 2 tables. *Journal of Statistical Software*, **42**(5), 1 – 23.

Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *The New Phytologist*, **11**, 37 – 50.

Jahoda, G. (2015). Quetelet and the emergence of the behavoiral sciences, *SpringerPlus*, **4**, 473 (10 pages).

Janson, S. & Vegelius, J. (1981). Measure of ecological association. *Oecologia*, **49**(3), 371 – 376.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementatation Using R*. Birkhäuser.

Kendall, M. G. (1952). George Udny Yule, C.B.E., F.R.S.. *Journal of the Royal Statistical Society, Series A (General)*, **115**(1), 156 – 161.

Kevles, D. J., Sturchio, J. L. & Carroll, P. T. (1980). The sciences in America, circa 1880. *Science*, **209**, 26 – 32.

Killion, R. A. & Zahn, D. A. (1976). A bibliography of contingency table literature: 1900 to 1974. *International Statistical Review*, **44**(1), 71 – 112.

King, G. (1997). *A Solution to the Ecological Inference Problem*. Princeton University Press.

King, G. (2004). EI: A program for ecological inference. *Journal of Statistical Software*, **11**(7), 1 – 41.

King, G., Rosen, O. & Tanner, M. A. (eds) (2004). *Ecological Inference: New Methodological Strategies*. Cambridge University Press.

Knudson, K. C., Schoenbach, G. & Becker, A. (2021). PyEI: A Python package for ecological inference. *Journal of Open Source Software*, **6**(64), 3397.

Kousser, J. M. (2001). Ecological inference from Goodman to King. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, **34**(3), 101 – 126.

Lancaster, H, O. (1966). Forerunners of the Pearson χ^2. *The Australian Journal of Statistics*, **8**(3), 117 – 126.

Lau, O., Moore, R. T. & Kellermann (2007). eiPack: $\mathrm{R} \times \mathrm{C}$ ecological inference and higher-dimension data analysis. *R News*, **7**(2), 43 – 47.

Le, C. T. (1998). *Applied Categorical Data Analysis*. Wiley.

Le, C. T. (2009). *Applied Categorical Data Analysis and Translational Research*. Wiley.

Leonard, T. (2000). *A Course in Categorical Data Analysis*. Chapman & Hall/CRC.

Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*. Wiley.

Lombardo, R. & Beh, E. J. (2016), The prediction index of aggregate data, *Journal of Applied Statistics*, **43**(11), 1998 – 2018.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719 – 748.

Meyer, D., Zeileis, A. & Hornik, K (2008). Visualizing contingency tables. In *Handbook of Data Visualization* (Chen, C., Härdle, W. & Unwin, A., eds), pp. 589 – 616, Springer.

Mirkin, B. (2001). Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, **55**(2), 111 – 120.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, **63**, 1 – 28.

Murphy, A. H. (1996). The Finley affair: A signal event in the history of forecast verification. *Weather and Forecasting*, **11**(1), 3 – 20.

Nightingale, F. (1858a). *Mortality of the British Army at Home, at Home and Abroad, and During the Russian War*. Harrison and Sons, London.

Nightingale, F. (1858b). *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army*. Harrison and Sons, London.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press, Toronto.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Lawrence Elbaum Associates, Hillsdale, NJ.

Nishisato, S. (2007). *Multidimensional Nonlinear Descriptive Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **5**, 157 – 175.

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation*. Drapers Memoirs. Biometric Series, Vol 1, London.

Pearson, K. (1930). *The Life, Letters and Labours of Francis Galton, Volume III: Correlation, Personal Identification and Eugenics*. Cambridge University Press.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, **4**, 453 – 454.

Plackett, R. L. (1974). *The Analysis of Categorical Data*. Charles Griffin & Company.

Powers, D. A. & Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis*. Emerald.

Quetelet, A. (1835). *Sur L'Homme et le Développment de ses Facultés: Essai de Physique Sociale.* Bachelier, Imprimeur-Libraire, Paris.

Quetelet, A. (1842). *A Treatise on Man and the Development of His Faculties.* William and Robert Chambers, Edinburgh.

Rayner, J. C. W. & Best, D. J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Chapman & Hall/CRC

Richardson, J. T. E. (1994). The analysis of 2 × 1 and 2 × 2 contingency tables: An historical review. *Statistical Methods in Medical Research*, **3**(2), 107 – 133.

Rovine, M. J. & Anderson, D. R. (2004). Peirce and Bowditch: An American contribution to correlation and regression. *The American Statistician*, **58**(3), 232 – 236.

Rowland, H. A. (1883). A plea for pure science. *Science*, **2**, 242 – 250.

Rudas, T. (1998). *Odds Ratios in the Analysis of Contingency Tables*. Sage University Press.

Rudas, T. (2018). *Lectures on Categorical Data Analysis*. Springer.

Schaeffer, J. T. (1986). Severe thunderstorm forecasting: A historical perspective, *Weather and Forecasting*, **1**(3), 164 – 189.

Schaeffer, J. T. (1990). The critical success index as an indicator of warning skill. *Weather and Forecasting*, **5**(4), 570 – 575.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer.

Sneath, P. H. A. (1957). Some thoughts on bacterial classification. *Microbiology*, **17**(1), 184 – 200.

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, **4**(2), 73 – 86.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press.

Stigler, S. M. (1995). Galton and identification by fingerprints. *Genetics*, **140**, 857 – 860.

Stigler, S. (2002). The missing early history of contingency tables. *Annales de la Faculté de Sciences de Toulouse Series 6*, **11**(4), 563 – 573.

Stokes, M. E., Davis, C. S. & Koch, G. G. (2003). *Categorical Data Analysis Using the SAS System*. Wiley/SAS Institute.

Sutradhar, B. C. (2014). *Longitudinal Categorical Data Analysis*. Springer.

Tang, W., He, H. & Tu, X. M. (2012). *Applied Categorical and Count Data Analysis*. CRC Press.

Tutz, G. (2011). *Regression for Categorical Data*. Indigo.

Upton, G. J. G. (2017). *Categorical Data Analysis by Example*. Wiley.

Van de Geer, J. P. (1993a). *Multivariate Analysis of Categorical Data: Theory*. Sage Publications

Van de Geer, J. P. (1993b). *Multivariate Analysis of Categorical Data: Applications*. Sage Publications

Van der Ark, L., Croon, M. A. & Sijtsma, K. (eds) (2005), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Lawrence Elbaum Associates.

Wakefield, J. (2004). Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society, Series A*, **167**(3), 385 – 445.

Wickens, T. D. (1989), Multiway Contingency Table Analysis for the Social Sciences. Lawrence Elbaum Associates.

Wrigley, N. (1985). *Categorical Data Analysis for Geographers and Environmental Scientists*. The Blackburn Press.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2. *Supplement to the Journal of the Royal Statistical Society*, **1**(2), 217 – 235.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, **3**(1), 32 – 35.

Yule, G. U. (1900). On the association of attributes in statistics: With illustrations from the material of childhood society. *Philosophical Transactions of the Royal Society of London, Series A*, **194**, 257 – 319.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, **2**(2), 121 – 134.

Yule, G. U. & Filon, L. N. G. (1936). Karl Pearson 1857 – 1936. *Obituary Notices of Fellows of the Royal Society*, **2**(5), 72 – 110.