NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J.; Smith, Derek R. “Real world occupational epidemiology, part 2: a visual interpretation of statistical significance” Archives of Environmental & Occupational Health Vol. 66, Issue 4, p. 245-248 (2011)

Available from: http://dx.doi.org/10.1080/19338244.2011.564235

Accessed from: http://hdl.handle.net/1959.13/1053775

Statistical Methods / Literature of EOH

Real World Occupational Epidemiology, Part 2: A Visual Interpretation of Statistical Significance

Eric J. Beh, BMath (Hons), PhD.

School of Mathematical & Physical Sciences, Faculty of Science and Technology, University of Newcastle, Callaghan, Australia

Derek R. Smith, PhD, DrMedSc, MPH.

School of Health Sciences, Faculty of Health, University of Newcastle, Ourimbah, Australia

Corresponding author:

Associate Professor Eric J. Beh

School of Mathematical and Physical Sciences,

University Drive, Callaghan, New South Wales 2308, Australia

Phone: +61 2 4921 5113

Fax: +61 2 4921 6898

Email: eric.beh@newcastle.edu.au

Introduction

Tests of significance in contingency tables represent one of the most useful techniques in occupational epidemiology, and one from which *Odds Ratios* (OR) can be derived to illustrate the probability that an event occurs compared to the probability that it does not. OR have become a regularly used tool for estimating the relationship between two variables, as well as offering a convenient interpretation of case-control studies.[1] OR are often used to signify another fundamental concept in *Environmental and Occupational Health* (EOH), that being the *Relative Risk* (RR) of contracting a particular disease following exposure to a particular hazard. RR estimated by the OR have become a de facto standard for representing 'hazard' in modern EOH,[2] partly out of tradition, and partly because the OR can provide a reasonable approximation of the RR given certain conditions.[3] In a previous paper,[4] we described the OR and RR, and reviewed many of their statistical and practical aspects by examining some of Irving Selikoff's classic asbestos research from the 1960s. There were a few reasons for choosing asbestos data in this regard.

Firstly, asbestos is the very model of an occupational hazard.[5] It represents one of the most heavily studied EOH hazards of human history[5] and one whose legacy still remains today in places such as Wittenoom, Western Australia[6] and Libby, Montana.[7] At an international level and despite repeated calls for a universal ban, asbestos is still with us.[8] Secondly, from an epidemiological perspective, asbestos represents one of the most important and historically significant case studies in EOH;[9] and one that regularly appears in, and is regularly cited by, its published literature.[10] Thirdly, Selikoff's original asbestos research was pioneering in nature and decisive in outcome, being one of the first studies to conclusively demonstrate

the dangers of working with this fibre and the magnitude of its risk. Furthermore, examination of Selikoff's original data is also very useful for highlighting various issues involved with the calculation and interpretation of workplace exposure data in hazard research. This is important because occupational epidemiology is constantly evolving as new statistical techniques emerge and the complexities of workplace exposures increase.[11]

As demonstrated in our previous article,[4] the OR provides one way to analyse an association between rows and columns, although its use is largely confined to the analysis of 2x2 contingency tables. In the current article, we explore other measures of examining associations where a contingency table consists of more than two rows and two columns. We shall demonstrate the role of the chi-squared test of independence, as well as correlation and correspondence analysis of data adapted from Selikoff's classic research into asbestos exposure and subsequent disease. Although his main study began in the early 1960s, some interesting raw data was later published in a 1981 issue of the *Bulletin of the New York Academy of Medicine*.[12] Parts of this data are reanalysed in our current article for demonstrative purposes (used with permission). Although correspondence analysis has been largely perceived as a descriptive, graphical means for understanding the statistical nature of relationships between rows and columns, it is actually based on complicated mathematical theories. Such discussions are beyond the scope of this paper, and therefore, we shall instead be focussing on the practical aspects of correspondence analysis and its application to the field of occupational epidemiology.

Selikoff's Original Asbestos Study

Irving J. Selikoff (1915-1992) was an American chest physician and pioneering researcher who has often been described as the country's foremost medical expert on asbestos-related diseases between the 1960s and the early 1990s.[13] In the late 1950s he opened a lung clinic in New Jersey and encountered a series of unusual illnesses among workers from a local asbestos plant.[14] In 1963, Selikoff collected data from a sample of around 1200 insulation workers in metropolitan New York. Clinical examinations were conducted to establish a diagnosis of asbestosis (and if so, its severity), while the period of occupational exposure to asbestos (if any) was also ascertained for each individual. Interestingly, most workers who had been exposed to asbestos for less than 20 years displayed normal chest films. Among those with 20 years (or more) exposure however, most chest x-rays were abnormal, and in many cases, extensively so. This 20-year time lag between exposure and disease became known as the '20-year rule'.[12]

From a statistical perspective, it would be of great benefit to determine the extent to which asbestos exposure is associated with the grade of asbestosis subsequently diagnosed in a worker, if any. To determine the nature of any such association, we shall simply refer to *Exposure* as the number of years that a worker has been exposed to asbestos. Similarly, *Grade* shall be used to reflect the grade of asbestosis that a worker has been subsequently diagnosed with. Table 1 summarises Selikoff's original data including the period of exposure classified according to five responses: 0-9 years, 10-19 years, 20-29 years, 30-39 years and 40+ years. Multiple grades of asbestosis are defined as either none, or ranging from Grade 1 (the least severe) to Grade 3 (the most severe). To illustrate our analysis of

the association between occupational exposure and disease severity, we shall begin by simply considering proportions. Note that for our discussion we shall not be inferring that a particular level of exposure will necessarily *cause* a particular grade, as an investigator should never assume from the beginning that association means causation.

Measuring Association in Selikoff's Original Data

Suppose we consider a worker who has been exposed to asbestos for less than 10 years. By calculation, the probability that a worker will not contract asbestosis is very high: 310 / 346 = 0.896, while for such a person, the probability they will contract Grade 3 asbestosis is 0. On the other hand, suppose we consider a worker who has been exposed to asbestos for at least 40 years. Then the probability they will not contract asbestosis is low, 7 / 121 = 0.058, while the probability that they will have Grade 2 or Grade 3 asbestosis is rather high: (51+28) / 121 = 0.653. While such simple summaries provide an indication of specific exposure / grade association, they do not provide a comprehensive insight into the global structure of this association. As such, a more formal approach for determining the strength of the association between exposure and grade is to determine the relative magnitude of the chi-squared statistic. A statistically large value suggests evidence of (in the sample analysed) a statistically-significant association between exposure and grade. Conversely, a statistically small (but positive) chi-squared value will mean that, based on evidence from the sample analysed, no such association exists.

For Table 1, the chi-squared statistic can be calculated as 648.8115, and with a p-value < 0.0001, indicates a statistically significant association between a worker's

exposure to asbestos and the grade of asbestosis they were subsequently diagnosed with. Again, caution must be exercised to ensure that one does not conclude that ‘a worker’s exposure to asbestos definitely causes the severity of asbestosis he / she is eventually diagnosed with’. While this may be demonstrated at a later time with more extensive research (and indeed this is what eventually happened with asbestos), an investigator should not make such assumptions in the beginning. In the aforementioned example, a chi-squared test of this type does not elicit such a uni-directional association structure. The calculated statistic does not determine the direction or nature of the association - only that such a statistical association exists.

To determine the direction of the relationship one may calculate the Pearson product moment correlation. In this calculation, a correlation lying between 0 and 1 indicates a positive association between two categorical variables, while a correlation between -1 and 0 indicates a negative association. When the rows and columns of a contingency table are not associated, then the correlation is zero. For Table 1, the correlation between exposure and grade is 0.69 and, with a p-value < 0.0001, indicates a statistically-significant positive correlation between exposure to asbestos and grade of asbestosis. That is, more severe cases of asbestosis are associated with more lengthy exposures to asbestos. However, the correlation does not elucidate at what point between the lowest level (Grade 1) or highest level (Grade 3) of exposure, that asbestosis will be contracted. To further understand the potential association between two variables, we can graphically examine its structure using correspondence analysis.[15, 16] As such, the following section will examine exposure

to asbestos versus the severity of asbestosis in a graphical format by way of correspondence analysis.

Graphical Depiction of Statistical Associations

Correspondence analysis is a technique that graphically displays row and column categories and allows for a visual comparison of their 'correspondences', or associations, at a category level.[15] For a practical guide to this topic consider, for example, Clausen,[17] Weller & Romney,[18] or Greenacre.[16] Recall that the chi-squared test of independence calculated from Table 1 revealed a statistically significant association between exposure and grade. A graphical representation of this association may be obtained by performing simple correspondence analysis. It is important to note that the term 'simple' does not necessarily refer to the simplistic manner in which the analysis is performed, as the mathematics involved is rather complex. Rather, the term 'simple' refers to the fact that it is the simplest of contingency tables (consisting of only rows and columns) that can be analysed. Figure 1 displays a two-dimensional plot of association between row and column categories from data in Table 1 and is referred to as a *correspondence plot*. Such a figure represents an important component of the output generated from a classical correspondence analysis of data in Table 1.

It reveals that, in general, a worker who has not been diagnosed with asbestosis is associated with a worker who has been exposed to asbestos for no more than 19 years. It also reveals an association between workers who have been diagnosed with Grade 2 or Grade 3 asbestosis (the two most severe grades of this disease) and those exposed to asbestos for 40 years or more. Furthermore, from the output, we

find that the first dimension visualises 84.2% of the association (as described by the chi-squared statistic) between the row and column categories, while the second axis visualises 15.4%. Thus, Figure 1 visually displays 99.6% of the association that exists between exposure and grade among the sample of New York asbestos workers. As a result, this correspondence plot provides an exceptionally good visual summary of the association between years of exposure to asbestos and a worker's subsequently diagnosed level of asbestosis.

Depending on how large a contingency table one is analysing, there will occasionally be a need to include more than just the first or second dimensions. The maximum number of dimensions needed to visualise any association will be: min(rows, columns) - 1. Therefore, the maximum number of dimensions needed to visualise the association between exposure and grade is: min(5, 4) - 1 = 3. The most conceptually difficult issue is being able to adequately visualise an association when more than three dimensions are required to summarise the data. This relates to the inherent difficulty in visualising anything requiring more than three dimensions. While there are certain tools that one may use to graphically depict associations in contingency tables,[19, 20] further discussion on this particular matter is beyond the scope of the current paper. Perhaps the most interesting finding revealed in Figure 1 is a visual confirmation of Selikoff's aforementioned '20 year rule'. Our graphical depiction clearly demonstrates that workers who have not been diagnosed with asbestosis are associated with either '0-9' or '10-19' years of exposure. Figure 1 also reveals some evidence (based on the data summarised in Table 1) to suggest that the 'mildest' form of asbestosis (Grade 1) is likely to commence at somewhere between 10 and 30 years of exposure.

Conclusion

In the current article we have utilised correspondence analysis to study some of Selikoff's pioneering asbestosis data from the 1960s. This technique is commonly used as a means of graphically summarising the association between variables under study. Even so, it is important to remember that correspondence analysis only considers whether there is a global association structure between the variables. One cannot simply say from the resulting correspondence plot that 'row response' *causes* 'column response', since causation is a very different concept to association. What can be inferred from the plot, however, is that the row responses are statistically associated with column responses. One aspect of correspondence analysis that has not been examined in the current article is the issue of multiple categorical variables. In these cases, multiple correspondence analysis can be used to obtain a graphical summary of the 'global' association. This matter will be examined in a future article.

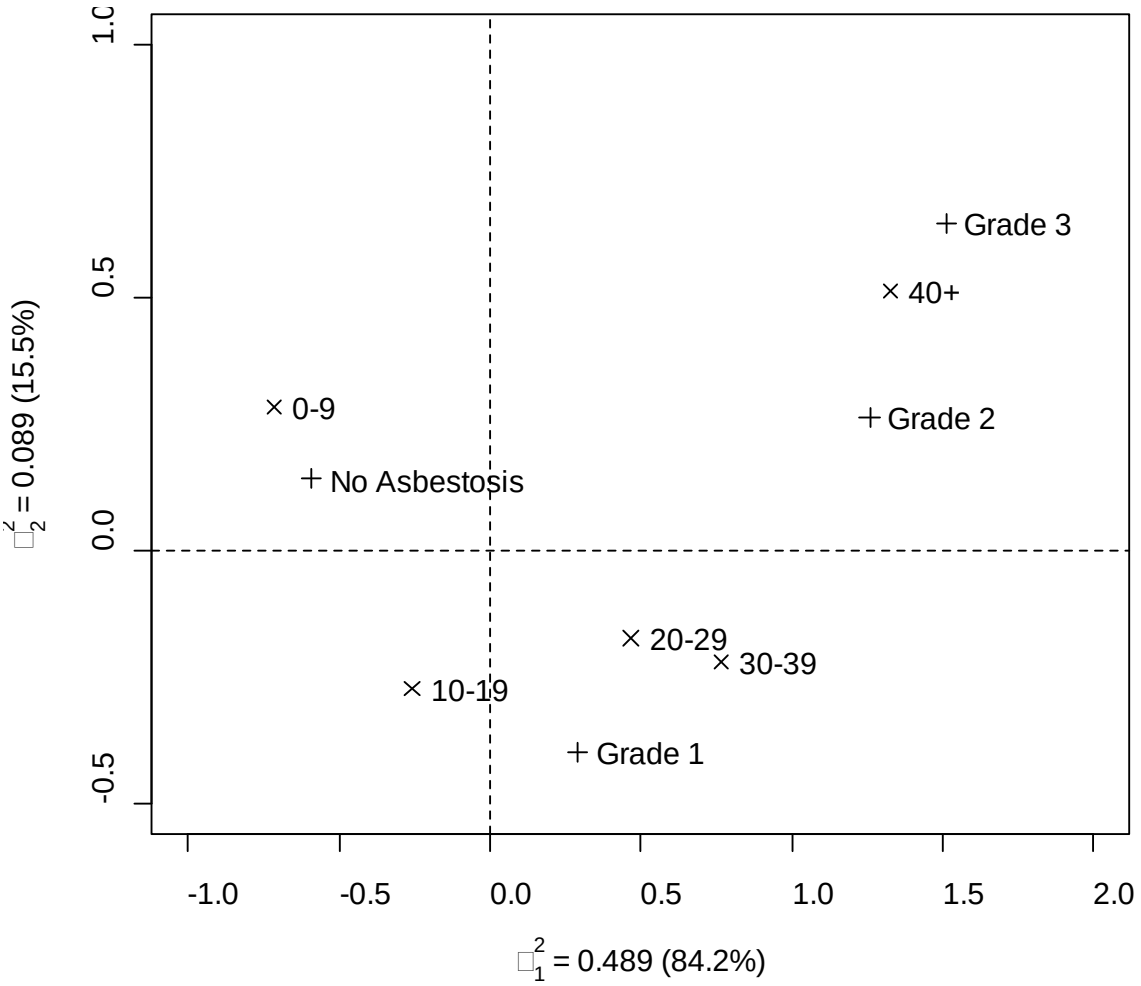
Table 1 Contingency Table Based on Selikoff's Original Asbestosis Data*

	Asbestosis Grade Diagnosed				
Exposure (years)**	**None**	**Grade 1**	**Grade 2**	**Grade 3**	**Total**
0-9	310	36	0	0	346
10-19	212	158	9	0	379

20-29	21	35	17	4	77
30-39	25	102	49	18	194
40+	7	35	51	28	121
Total	575	366	126	50	1117

*Adapted from Selikoff (1981)[12] published in the *Bulletin of the New York Academy of Medicine* (used with permission), **Years of occupational exposure to asbestos

Figure 1 Correspondence Plot Derived from Selikoff's Asbestosis Data*



*Derived from data by Selikoff (1981)[12] published in the *Bulletin of the New York Academy of Medicine* (used with permission)

References

1. Bland JM, Altman DG. Statistics notes. The odds ratio. *BMJ.* 2000;320:1468.

2. Smith DR, Attia J, McEvoy M. Exploring new frontiers in occupational epidemiology: the Hunter Community Study (HCS) from Australia. *Ind Health.* 2010;48:244-8.

3. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA.* 1998;280:1690-1.

4. Beh EJ, Smith DR. Real world occupational epidemiology, Part 1: Odds ratios, relative risk and asbestosis. *Arch Environ Health.* 2011;66:(in press).

5. Guidotti TL. Why study asbestos? *Arch Environ Occup Health.* 2008;63:99-100.

6. Smith DR, Leggat PA. 24 years of pneumoconiosis mortality surveillance in Australia. *J Occup Health.* 2006;48:309-13.

7. Bandli BR, Gunter ME. A review of scientific literature examining the mining history, geology, mineralogy, and amphibole asbestos health effects of the Rainy Creek igneous complex, Libby, Montana, USA. *Inhal Toxicol.* 2006;18:949-62.

8. Collegium Ramazzini. Asbestos is still with us: repeat call for a universal ban. *Arch Environ Occup Health.* 2010;65:121-6.

9. Smith DR, Beh EJ. Occupational epidemiology in the real world: Irving Selikoff, odds ratios and asbestosis. *Arch Environ Health.* 2011;66:(in press).

10. Smith DR. Highly cited articles in environmental and occupational health, 1919-1960. *Arch Environ Occup Health.* 2009;64 (Suppl.):32-42.

11. Guidotti TL. Occupational epidemiology. *Occup Med (Lond).* 2000;50:141-5.

12. Selikoff IJ. Household risks with inorganic fibers. *Bull N Y Acad Med.* 1981;57:947-61.

13. McCulloch J, Tweedale G. Science is not sufficient: Irving J. Selikoff and the asbestos tragedy. *New Solut.* 2007;17:293-310.

14. Smith DR. The historical development of academic journals in occupational medicine, 1901-2009. *Arch Environ Occup Health.* 2009;64 (Suppl.):8-17.

15. Beh EJ. Simple correspondence analysis: A bibliographic review. *Int Stat Rev.* 2004;72:257-284.

16. Greenacre MJ, *Correspondence Analysis in Practice (Second Edition)*. 2007, London: Chapman & Hall / CRC Press. 280pp.

17. Clausen SE, *Applied Correspondence Analysis: An Introduction*. Sage University Papers Series on Quantitative Applications in the Social Sciences. 1998, Thousand Oaks: Sage. 80pp.

18. Weller SC, Romney AK, *Metric Scaling: Correspondence Analysis*. Sage University Papers Series on Quantitative Applications in the Social Sciences. 1990, Thousand Oaks: Sage. 96pp.

19. Friendly M, *Visualizing Categorical Data*. 2000, Cary: SAS Institute Inc. 456pp.

20. Meyer D, Zeilis A, Hornik K, *Visualizing contingency tables*, in *Handbook of Data Visualization*, Chen C, Hardle W, and Unwin A, Editors. 2008, Springer: Berlin. p. 589-616.