OPINION

Talking Statistics: A reflection on some of the problems with statistical language

Lois Rollings, Learning Enhancement Team, Middlesex University, London, UK.
L.Rollings@mdx.ac.uk

Keywords: statistics, language.

1. Introduction

For most of my life I managed to swerve statistics. I learned a little at school and studied some (very theoretical) statistics as part of my mathematics degree. As a teacher, in schools and then university, I did not teach anything beyond GCSE statistics. It was not until I got a post in Mathematics and Statistics support a decade ago that I had to begin to learn the subject properly.

I had excellent support from the **sigma** network, attending a memorable SPSS Bootcamp and other events which helped me enormously. But I was conscious that there were various aspects of statistics that presented problems for me. One was the way in which statistics differed from mathematics in being much less cut and dried. If a student had made a mistake in a calculation or argument it was fairly easy to spot and correct. However, when a student said 'my supervisor said I should do a t-test' and this did not seem the most appropriate way forward it was much harder to advise. I also realised that I was finding the language of statistics harder to master than I felt it ought to have been. It is this aspect that I will focus on in this article as I tentatively suggest that students might also have such problems.

2. Statistical Language

Language acquisition is not a trivial process (Leung 2005, cited by Kaplan et al., 2010, p.1) and specialized language can lead students to perceive the subject as more difficult than it is (Lemke 1990, cited by Kaplan et al., 2010 p.2). The language of a discipline can be seen as "*a tool for participation in a community of practice*" (Espinet et al., 2012, cited by Kaplan and Rogness, 2018, p.1) and it is surely part of our job to induct students into that community.

2.1 Statistical English and Lexical Ambiguity

Rangecroft (2002) discusses how Mathematical English (ME) differs from Ordinary English (OE) and suggests that Statistical English (SE) is distinct from both. I considered myself fairly fluent in OE and ME but it was with SE that I had difficulties.

One issue is that statistics often uses words from OE but gives them a particular technical meaning. This can give rise to 'lexical ambiguity' (Barwell 2005, cited by Kaplan 2011, p.56). An everyday example of this is that if you are reading a sentence starts 'The bat was…' and this brings to mind a cricket or baseball bat but then the sentence continues 'roosting in the belfry', it will take a moment for you to change your perspective. It has been shown that this disambiguation requires higher levels of brain activation than reading a sentence where there is no ambiguity (Mason and Just, 2007, p.115).

There are many terms used in SE that have a meaning in OE but where the statistical meaning differs from the everyday meaning. In some cases the two meanings are close; in others the difference can be quite large. Some examples are 'significant', 'random' and 'spread' and you can doubtless think of others. This means that students might start to think of the everyday meaning and might take a while

to readjust to the statistical meaning. Perhaps a more serious concern is that they might not realise that the technical meaning is different and so persist with the everyday meaning. There is also the added complication that words used in both mathematics and statistics, for instance 'normal' and 'estimate', may not have the same meaning in both contexts.

Konold 1995 (cited by Kaplan et al., 2010, p.2) reports that statistics students have "*strongly-held, but incorrect, intuitions that are highly resistant to change*". This brings to mind the Piagetian ideas of 'assimilation' and 'accommodation'; it is much easier to assimilate an idea into an existing schema than to have to reconfigure the schema (Mason and Johnston-Wilder, 2004, pp.149-151). The problem is of course compounded if as teachers we do not realise that what we are talking about is not what is being understood by the students, and this issue may occur in other subjects. This means that "*there is a vast amount of room for misunderstanding that may never be detected*" (Roth, 2005, cited by Kaplan and Rogness, 2018, p.1).

The term 'random' is a good example of this. Perhaps particularly with younger students the everyday meaning of random is 'weird'. Kaplan et al. (2014, p.12) found that "*students persist in thinking that random processes are haphazard, weird or have unlikely outcomes*" and describe using two images in their teaching to try to address this. A picture of some people dressed as zebras illustrated the everyday meaning and a hat illustrated the technical statistical meaning. They report some success in using this strategy as throughout the statistics course they could remind students that they meant the "*hat not the zebras"* (Kaplan et al., 2014, p.13).

In some case it is the formation of a 'nominal group' that provides a technical meaning, as in 'standard deviation' and 'degrees of freedom'. Some of these groups are 'dense', for instance 'Pearson's product-moment correlation coefficient r'. Different selections of these words can be found in different sources and exactly what is meant can differ. From a linguistics perspective the question is which of these is the 'head noun'. Are we talking about the abstract idea of correlation or the coefficient? As McConway (2016) remarks "*product has nothing to do with the production of goods and moment has nothing to do with time*".

The issue of lexical ambiguity is not confined to mathematics and statistics. Rector et al., (2013). discuss its presence in scientific disciplines and evolutionary biology in particular. It is also not peculiar to English; Ilana Lavy & Michal Mashiach-Eizenberg (2009) describe similar issues in Hebrew. As well as there being an overlap between OE and SE, there may also be an overlap with technical terms used in other disciplines. In accounting 'variance' just means difference and I have been confused by this myself. 'Sample' might have a different meaning for biomedical students (Dunn et al., 2016, p.14). It might seem counterintuitive but lexical ambiguity might be less of a problem for students whose first language is not English as they may not be so familiar with the everyday meanings of words (Dunn et al., 2016, p.10).

Another potential problem is that statistical language is not standardised (Dunn et al., 2016, pp.9-10). Variables might be described as 'predictor', 'explanatory' or 'independent'. Different software packages also use different terminology; Minitab refers to a '2-sample' t-test whereas SPSS uses 'independent sample' t-test. This can be a particular issue for those of us working in a support context as we will not always know which term the student is used to. The same word can also be used with different meanings. For instance, 'average' might refer to the mean (as it does in Excel) or as a term for one of several averages (mean, median or mode). An average might also be referred to as a 'measure of centre' or a 'measure of central tendency' and students might not realise that these are all the same thing. Even in the name of the discipline itself there is a distinction between the study of 'statistics' and a 'statistic'.

2.2 Statistician's Names and Definitions

Something I found particularly confusing at first was the profusion of statistician's names attached to different tests or calculations. For some reason it took me a long time to distinguish the Mann-Whitney and Kruskal-Wallis tests. (And embarrassingly much longer to discover that in each case the names refer to two people not one with a double-barrelled surname.) In some SPSS output there will be several options (all named) for a given test and differing opinions on which one should be used. In some ways I should be used to this as many mathematics theorems are 'named'. But although the same theorem might have different names attached to it perhaps out of national pride (see the inequality named after, variously, Cauchy, Schwarz and Bunyakovskii) there will not be alternative versions of the same theorem. This highlights one of the key differences between mathematics and statistics; once a mathematical theorem is accepted as true, it will not be challenged, whereas in statistics there might be no standard way of doing a calculation.

One thing which still strikes me as odd is that it seems very rare for sources to explain who these people were. Of course, there are other issues, beyond the scope of this article, in that several statisticians of the early twentieth century were eugenicists. This has been acknowledged by some institutions, for instance University College London has 'denamed' buildings (UCL, 2020), but not, as far as I know, in statistics textbooks.

Although many statisticians' names appear, statistical writing can seem impersonal. Use of the passive "*obscures the human agency involved"* (Morgan 2005, p.110) and increases 'social distance'. (This was a linguistic term long before it was used in relation to a pandemic.) I suggest that this might be a particular problem for students from social sciences. As Graves et al. (2014, p.7) remark *"[definition] performs different roles among disciplines"*. In social sciences and lower level maths a definition might formalise a pre-existing concept. An author might cite several examples and then say 'We call this X'. In mathematics, certainly at university level, it is usual to start with a definition, 'X is defined as…' and this formulation is often used in statistics. In statistics definitions might not even always be agreed (Dunn et al., 2016, p.17).

3. Concluding Remarks

A key question is what we can do about these issues but, just as there are different ways of calculating, say, an effect size, there is no consensus on the approach that we should adopt.

Rangecroft (2002, p.36) encourages "*trying to pre-empt difficulties by careful choice of language*" and suggests that we should "*become more attuned to the possibilities of misunderstandings arising from language difficulties*". Kaplan and Rogness (2018, p.4) concur, saying that one of the first steps is for "*instructors to recognise words that have lexical ambiguity*".

Some argue for a more consistent use of technical terms. Kaplan et al. (2011, pp.59-60) advocate using the term 'variability' rather than 'spread' but warn not to expect "*a change in terminology to be a panacea*". Kaplan et al. (2010 p.17) suggest using 'measures of centre' instead of 'averages' but, as Dunn et al. (2016, p.20) remark, it would be "*almost impossible*" to get consistency across sources and we need to respect different usage in different disciplines.

While Lavy and Mashiach-Eizenberg (2009, p.7) argue that technical terms should be introduced using both formal and informal language, others, for instance Neibert et al. (2012, cited by Kaplan and Rogness, 2018, p.3), say that using common everyday words as an alternative to technical terms might increase misunderstandings and barriers for students. Dunn et al. (2016, pp.20-24) discuss at some length the different solutions offered by others and conclude that a balance needs to be struck

"*between the mathematical and linguistic requirements"* of a statistics course, noting that how this is achieved will depend on the context.

Finally, it is perhaps worth noting that our language use, particularly when speaking, is not conscious (Martin, 2010, p.14). So, even if we could agree on an approach, we will not always get this right.

4. References

Dunn, P.K., Carey, M.D., Richardson, A.M. and McDonald, C. (2016). Learning the language of statistics: Challenges and Teaching Approaches. *Statistics Education Research Journal,* 15(1), pp.8-27. Available at: http://iase-web.org/Publications.php?p=SERJ [Accessed 7 November 2023].

Graves, H., Moghaddasi, S. and Hashim, A. (2014). "Let *G=(V, E)* be a graph": Turning the abstract into the tangible in introductions in mathematics research articles. *English for Specific Purposes,* 36, pp.1-11. https://doi.org/10.1016/j.esp.2014.03.004

Kaplan, J., Rogness, N.T. and Fisher, D.G. (2010). Lexical Ambiguity in Statistics: How Students Use and Define the words: Association, Average, Confidence, Random and Spread. *Journal of Statistics Education,* 18(2), pp.1-17. https://doi.org/10.1080/10691898.2010.11889491

Kaplan, J., Rogness, N.T. and Fisher, D.G. (2011). Lexical Ambiguity: making the case against spread. *Teaching Statistics,* 34(2), pp.56-60. https://doi.org/10.1111/j.1467-9639.2011.00477.x

Kaplan, J., Rogness, N.T. and Fisher, D.G. (2014). Exploiting lexical ambiguity to help students understand the meaning of *random. Statistics Education Research Journal,* 13(1), pp.9-24. Available at: http://iase-web.org/Publications.php?p=SERJ [Accessed 7 November 2023].

Kaplan, J. and Rogness, N.T. (2018). Increasing Statistical Literacy by Exploiting Lexical Ambiguity of Technical Terms. *Numeracy,* 11(1), Article 3. https://doi.org/10.5038/1936-4660.11.1.3

Lavy, L. and Mashiach-Eizenberg, M. (2009). The Interplay Between Spoken Language and Informal Definitions of Statistical Concepts. *Journal of Statistics Education,* 17(1). https://doi.org/10.1080/10691898.2009.11889502

Martin, J. (2010). Language, register and genre. In C. Coffin, T. Lillis and K. O'Halloran, eds. *Applied Linguistics Methods: a Reader*. London: Routledge in association with the Open University, pp.12-32.

Mason, J. and Johnston-Wilder, S. (2004). *Fundamental Constructs in Mathematics Education,* Abingdon: RoutledgeFarmer.

Mason, R. and Just, M.A. (2007). Lexical ambiguity in sentence comprehension. *Brain Research,* 1146, pp.115-127. https://doi.org/10.1016/j.brainres.2007.02.076

McConway, K. (2016). More confusing terms in statistics. *Open Learn*. Available at: https://www.open.edu/openlearn/course/view.php?id=12784 [Accessed 7 November 2023].

Morgan, C. (2005). Words, Definitions and Concepts in Discourses of Mathematics, Teaching and Learning. *Language and Education,* 19(2), pp.103-117. Available at: https://discovery.ucl.ac.uk/id/eprint/1564242/ [Accessed 7 November 2023].

Rangecroft, M. (2002). The Language of Statistics. *Teaching Statistics,* 24(2), pp.34-37. https://doi.org/10.1111/1467-9639.00080

Rector, M.A., Nehm, R.H. and Pearl, D. (2013). Learning the Language of Evolution: Lexical Ambiguity and Word Meaning in Student Explanations. *Research in Science Education*, 43, pp.1107-1133. https://doi.org/10.1007/s11165-012-9296-z

UCL (2020). *UCL denames buildings named after eugenicists*. Available at: https://www.ucl.ac.uk/news/2020/jun/ucl-denames-buildings-named-after-eugenicists [Accessed 1 November 2023].