Statistics for Multimodality: why, when, how — an invitation

John A. Bateman
Bremen University
bateman@uni-bremen.de

Tuomo Hiippala
University of Helsinki
tuomo.hiippala@helsinki.fi

Abstract

Although the number of articles in visual and multimodal communication that include statistical validation of claimed results is increasing, we suggest in this article that this is by no means enough. Statistical methods should belong to every multimodality researcher’s toolset precisely because the phenomena under study are subtle and complex. Without applying basic statistical techniques, it is difficult to know whether one ‘has’ a result or not. Moreover, even if statistical tests are not undertaken, bearing in mind that they should – in principle – be possible at some stage of research adds clarity to methodology and design. Improving awareness of this issue is particularly important at this time because multimodality now faces the challenge of moving on from a period of broad theory construction to studies that build on those foundations. Reliably locating significant patterns in data is a crucial precondition for such work and statistical methods provide established and well-understood techniques for just this task. The primary intended audience of this article are researchers, and particularly junior researchers, who have not so far embraced such methods. We seek to show that the application of statistical methods is nowadays relatively straightforward and that adding these techniques is of considerable value both for conducting effective research and for convincing others of that effectiveness. Conversely, within many contemporary research contexts, not applying such methods is simply unnecessarily disempowering.

1 Introduction and motivation

A growing number of researchers in visual studies and multimodality analyse bodies of selected data in their graduate and postgraduate research, and established multimodality scholars increasingly see the need to engage in larger scale studies when applying aspects of multimodality theory. Such work demands methods for finding patterns in any data gathered but this is currently still seen as a major unresolved challenge. One relatively common approach observed in the literature is to rely on visualisations – examples here include Manovich’s (2012) cultural analytics approach, Caple et al.’s (2018) recent proposals for diachronic visualisation, many techniques used in the digital humanities (see e.g. Jänicke et al. 2016), as well as many of the visualisations offered in, sometimes quite expensive, ‘computer-assisted qualitative data analysis software’ (CAQDAS). However, despite an intuitive appeal, relying on visual perception for finding patterns only works when the data exhibits relatively straightforward and superficial patterns. This can certainly be useful but is intrinsically limited. Raw data in general is unlikely to reveal interesting patterns because human perceptual abilities, although powerful, only deal with clear patterns of surface contrasts, such as perceptible changes in contrast or rhythm. Many significant inter-relationships and dependencies are not so directly accessible without looking beyond the surface regularities in a variety of ways. In contrast, obtaining reliable, and above all predictive, views of data that *can* reveal patterns is one of the things that statistical approaches to data analysis are precisely designed to support. Being able to apply such techniques for the purposes of multimodality research is therefore now of considerable importance.

This step is relevant for the majority of issues discussed for visual communication and multimodality research. For example, several introductory text books related to visual analysis, such as Ledin and Machin (2018), round up their chapters with potential research projects for students which *can all be usefully complemented by statistically evaluating any results obtained*. The typical ‘empirical’ research design encouraged in such projects runs as follows: select some examples for discussion, develop or adopt a classification framework for assigning the data to categories, and then talk about the resulting classifications as support for the points that one wanted to make about the data in the first place. The typical ‘theoretical’ research design is similar: first, some examples are selected; second, a classification scheme is suggested arguing discursively for any divisions between categories by drawing on examples that appear to support or require the division under discussion; and third, the resulting classification

scheme is claimed to be the way in which data of the kind analyzed is organized (cf. Kress and van Leeuwen 2006). Unfortunately, as Thomas (2014) argues, approaches of these kinds, now repeated in many dissertations and articles, can often end up as being circular. Without statistics, we cannot know whether any features identified represent simply random variation or exhibit meaningful patterns, i.e., patterns that can function predictively with respect to the materials addressed. In short, such studies do not engage with the data in ways which could show that they are wrong and fail to perform the most basic of checks that they are producing usable results. Building further research on such ‘results’ is consequently a precarious business indeed and provides an unnecessarily weak basis for multimodality research; interesting questions may be raised but methodologies appropriate for tackling those questions are marginalized.

Statistical analyses are sometimes not performed because of an assumed difference in orientation between more qualitatively-oriented, discursively-constructed work and more quantitative approaches. Within multimodality research construed broadly, such divisions align with delineations between work that has traditionally drawn little on more formal, quantitative methods and those disciplines where some kind of statistical knowledge is largely taken for granted, ranging across psychology (including psycholinguistics and visual perception), corpus linguistics (Baayen 2008; Gries 2016), face-to-face interaction studies (Allwood et al. 2007), and content analysis (Krippendorff 2004). Here, we reject distinctions between so-called qualitative and quantitative research and promote what are generally nowadays termed ‘mixed methods’ paradigms throughout (cf., e.g., Teddlie and Tashakkori 2011; Schoonenboom and Johnson 2017). Although being able to ‘count’ or ‘measure’ occurrences of some selected properties is necessary for applying statistical techniques, this by no means constitutes a restriction to quantitative research. Essentially, we are always looking for patterns and those patterns should have qualitative import. At the same time, patterns are necessarily *re-occurrences* and a re-occurrence is inherently something that allows quantification. Re-occurrences of interest can then be across ‘texts’, within ‘texts’, across time, across cultures, across individuals, across forms of expression, all of which interface directly with qualitative research questions. This undercuts any notion that statistical methods are somehow something ‘foreign’ to qualitative research and are only imposed on qualitative data by methods more appropriate for the natural sciences. The search for meaningful patterns transcends such boundaries. The more data that one can consider in this enterprise, the more reliable (i.e., predictive) and representative such patterns can become. But, at the same time, the more data one has, the more difficult it is to find patterns by inspection. One of the most powerful ways of finding patterns currently available is precisely that provided by statistical methods and so it is scarcely logical to avoid applying such techniques to multimodality research, even though the patterns we will be looking for will always be qualitatively, i.e., interpretatively, significant patterns.

As a consequence, we consider the broad situation currently prevalent in many branches of multimodality research (which we take to include visual communication), in which work offering statistical validation of claims remains comparatively rare, to be inappropriate. Not performing statistical evaluations of empirical research reflects a tradition that is no longer functional. It increasingly cuts off researchers from presenting their results to fields (and even sub-fields) beyond their own local academic contexts and is thereby increasingly disempowering. Moreover, it makes it possible on the one hand to make claims about data that might actually not hold and, on the other hand, and arguably worse, it offers support for presentation strategies which blur crucial methodological differences between *claims*, or *hypotheses*, and *results*. Claiming that some data exhibits a pattern is not yet a predictive result: it is a hypothesis of a pattern that may hold. Although this can itself be of enormous value, precisely because of the questions that it may generate, it is nevertheless crucial that such claims subsequently be subjected to empirical validation so that they can be extended, improved and, if necessary, replaced (cf. Bateman 2019).

Our main purpose in this article is therefore to motivate multimodality researchers who are not yet familiar with statistical techniques to take some first steps towards applying basic statistical methods when exploring any patterns they believe to be present in their data. We provide this motivation primarily by showing that there is now little justification for not engaging with statistics; freely available tools make it straightforward to include basic (and not so basic) statistical operations within a research project as a matter of course. We also aim to give researchers some intuitions concerning just *why* the statistical techniques work; this intuition should then serve to help engage more productively with the much broader range of statistical techniques now becoming available. Spin-offs of such engagement include more effective ways of collecting and organising data as well as more sophisticated ways of presenting results. We show this using specific examples drawn from multimodality research, although the basic techniques are applicable to almost any piece of empirical research conducted in multimodality and

visual communication.

The article is organised as follows. First, we present the basic motivations and conceptualisations underlying the notion of ‘significance testing’ in statistics, still the most commonly reported kind of statistical evaluations found in the literature. Second, we introduce the computational tool that we recommend for all kinds of statistical work. Third, we run through case studies building on the basic notions introduced, showing how several statistical techniques are directly applicable to common tasks that always arise in multimodality research. Finally, we conclude with a brief review of what has been covered and offer explicit pointers for the most useful ‘next steps’ to be taken in order to engage more fully and usefully with statistical methods when carrying out multimodality research. We do not assume any previous mathematical or statistical knowledge and take considerable pains to make the discussion as straightforward as possible. We also are *not* arguing that all research must be empirical in orientation – as suggested above, finding revealing hypotheses is itself crucial for advancing research. What we are arguing, however, is that *if work is intending to be empirical*, then either statistical methods should be present or explicit reasons should be offered as to why the study or the data did not support the level of evaluation that statistical tests make possible. In fact, designing research that could, at least in principle, be subjected to statistical evaluation tends to produce better and more valuable research *even if the step of statistical evaluation is not then actually done* precisely because of the methodological clarity enforced.

2 The basic idea: everything varies but that variation is more often than not systematic

The essential idea behind applying statistical methods is to assess whether results that have been found are, indeed, results – i.e., whether any patterns or relationships claimed to be present may actually be used *reliably* to build arguments and predictions. Being able to offer support for such conclusions in ways that go beyond discursive argumentation is a crucial step, one that marks out increasing applicability in any area of study. Statistical methods offer a well understood and extensive set of tools for supporting this move and for showing when, and when not, it could be appropriate.

One of the most fundamental motivations for employing statistics for model building is then actually a very simple one: we can rarely look at *all the cases* of any data or phenomena that we are interested in. That is, we can only look at some selection – a ‘sample’. This rather obvious point has two rather important consequences:

1. It raises the question of how we can know that the data we *did* look at tells us something about how things work more generally. If what we happened to look at does not reflect, or is not representative, of the patterns being sought in the ‘broader picture’, then any conclusions we come to may be of much more limited validity than we were hoping.
2. Whenever we do look at some selection rather than the whole, the precise results we get *will vary depending on our selection*. Whenever something is ‘measured’ or ‘counted’, the numbers obtained will exhibit intrinsic *variation*, that is, the number of patterns found among some set of objects under investigation will differ. This point is critical because we need to know to what extent any differences we find are due to some real difference in the data we are exploring or is just due to this intrinsic, and unavoidable, ‘sampling variation’.

These two points – representativeness and intrinsic variation – explain and motivate most statistical techniques because those techniques share a common goal: they seek to match observed data with a description of hypothesised properties of the data as a whole. A rather more general, and powerful, way of construing this is to do so semiotically, because statistical models are intended to *describe* data. In fact, in a very real sense that we will use at various points as we proceed, statistical models are *metaphors* in relation to the ‘world’ of phenomena being described and *diagrams* with respect to the data: both terms applied with a strict Peircean semiotic reading (cf. Bateman 2018). Models are therefore ways of capturing regularities in a manner supportive of inference; they are built to draw out abstract properties inherent to a body of data (i.e., act as diagrams) which, if representative, may be used to make predictions about another domain, the world at large (i.e., act as metaphors). Statistics gives us methods supportive of both stages of investigation – drawing patterns out of data and evaluating those patterns’ reliability and generality. The former concerns are often called *descriptive statistics* and the latter *inferential statistics*, although as we have couched them here, both involve, or support ‘inferences’ – just inferences of different kinds. Within inferential statistics there is also a distinction increasingly drawn between accounts that perform ‘Null-Hypothesis Significance Testing’, or NHST, and those that adopt a Bayesian approach.

We will see what null hypothesis testing involves in a moment as this is still the most common kind of statistical testing reported in the literature. Much of our phrasing is, however, rather more sympathetic to the Bayesian approach; we return to this at the end of the paper when we re-consider modelling.

Measuring to what extent variation in some analyzed data is *significant* variation and not just a consequence of the fact that analyzed data will unavoidably vary by itself anyway is enabled by some interesting properties inherent to many natural phenomena. Even though the ‘sampling’ of selected measures or counts never leads to the ‘same’ values, what occurs instead is that results will *group themselves* subject to regularities that can be used to characterize, or ‘model’, the situation that one has. Statistics provides a range of standard techniques for gaining access to such regularities and using them for inference. To begin, therefore, we introduce some of the basic ideas by describing approaches that rely on ‘approximating’ the data to show regularities. Describing regularities in this way works by compressing raw information into more compact, summarising representations; i.e., to build diagrams in Peirce’s technical sense. One of the simplest such approximations is, for example, the average, or *mean*, of some set of values obtained by measurement. The average of a set of values is an approximate ‘model’ of the set of values as a whole, although semiotically it has clearly lost a considerable amount of information: if we use this as a predictive model, i.e., only selecting values that are equal to the average value, then it will often be wrong. But this is still a better model than picking values at random. We can then ask how to make the model better, i.e., more reliably predictive, in *a principled fashion*.

A far better approximating model can be obtained by additionally calculating how widely the values vary from the mean: that is, how ‘spread out’ they are. If we know both this and the mean, then we already have quite a good, if approximate, representation of the data. The usual way of characterising variation is to calculate what is called the *standard deviation* from the mean, that is, the ‘average’ extent to which the values diverge from the mean; we show how to do this below. Both mean and standard deviation relate to how frequently particular values occur in relation to others and together tell us a surprising amount. A fundamental (mathematical) result from statistics is that any measurements taken in a sample will tend to cluster around the actual mean value of the ‘population’ the sample is taken from and, moreover, that this clustering itself will have a particular ‘shape’. Finding out about this shape is then the first step towards understanding what most statistical measures are doing and why they work. In short, as soon as the generalized shape of the variation for some measure is known, we have a good model of the data that supports a range of useful inferences. In particular, given the mean and the standard deviation it is possible to calculate how *likely* particular values are: values close to the mean will be likely and values far from the mean will be unlikely. Moreover, just *how* unlikely depends on the standard deviation. If the standard deviation is large, then values far from the mean are respectively more likely; if the standard deviation is small, then values far from the mean are correspondingly less likely because the data is less widely ‘spread out’.

We can visualise the shape of such measurements by simply taking samples and counting up how often particular values occur. If put in a graph, we then see directly how frequently a particular value was found: the more often a particular value occurs, the higher the corresponding point in the graph will be, and the less often, the lower the corresponding point will be because its count is less. When such graphs are drawn, a common shape that emerges is the ‘bell curve’ of the so-called *normal distribution*, three examples of which are shown in Figure 1. Many phenomena follow this kind of curve: that is, simply by taking measurements, counting up how often particular values are found, and plotting these counts in a graph, the result will commonly be a curve shaped in this fashion. The three curves in the figure show the effects of differing standard deviations. Each individual normal distribution is in fact uniquely defined by the values of the mean (M) and standard deviation (SD): a low standard deviation means that the values do not spread out very much and so values will not deviate much from the mean. This gives a much steeper curve such as the one on the left in the figure. A higher value for the standard deviation, in turn, results in a flatter curve, as exemplified by the one in the middle.

Regardless of the precise shape of the bell curve, all three samples in Figure 1 follow a normal distribution, and this means that several purely mathematical properties for normal distributions in general apply. For example, in a normal distribution, 68% of values fall within one standard deviation on either side of the mean, while 95% fall within two standard deviations from the mean. This is then used for statistical ‘testing’ as follows. Note first that if something is more frequent, it is more likely to occur – i.e., something that occurs 60 times out of a hundred is more likely to occur than something that occurs 20 times out of a hundred. This means that the height of the curve at any point is *an indication of how likely that value is*. It is then a small step to establish that the probability of finding some value in a particular *range* of values is simply the *area under the curve* for that range of values; Bateman et al. (2017: 176–178) show why this is with a simple example. Combining this with the mathematical

Figure 1: Three sets of samples each following a normal distribution. Each sample differs in terms of their mean (M) and standard deviation (SD). The X-axis (horizontal) gives the values, whereas the Y-axis gives the probability of randomly drawing that value from the sample on the X-axis. The shape of the normal distribution depends on how far the values disperse from the mean, which is measured by standard deviation. Plot created using R 3.6.0.

properties just given means that a value drawn randomly from some normally distributed sample will fall within two standard deviations of the mean in *95 cases out of 100* (i.e., 95%). And conversely, if we find a value that is *further* than 2 standard deviations away from the mean, then the likelihood of that occurring is less than 5%, i.e., in just 5 cases out of a hundred, or quite rarely.

This is useful to know because it offers a relatively straightforward way of ascertaining how likely some state of affairs is – *given the assumption of a particular curve*, or ‘model’. Consider the cases shown in Figure 1 with three data sets. Imagine that we find some particular value, say 9, and wonder what set of data it fits to best. We can see directly that there is a very small chance that it is drawn from the rightmost curve because it is far more than 2 standard deviations away from the mean for that curve (in fact, it lies more than 6 standard deviations away, as we can see by multiplying by the standard deviation and subtracting from the mean: $28 - 6.33 \times 3 = 9$). Considering the other two curves, the situation is visually less clear; but, performing the same calculations, we see that the value 9 is 1.5 standard distributions away from the mean of the leftmost curve ($6 + 1.5 \times 2 = 9$) but only 0.75 standard deviations away from the mean of the middle curve ($12 - 0.75 \times 4 = 9$). This allows us not only to state that the value most likely comes from the middle data set but also to quantify precisely just *how much* more likely that is. In fact, using the area under the curves, we have a 45% chance of finding something 0.75 standard deviations or more from the mean, whereas for 1.5 standard deviations or more the chance goes down to 13%.

Many statistical tests then work in precisely this way: first, a particular shape of probability curve is matched to some data (for example, by working out the mean and standard deviation of samples and assuming a normal distribution). This is taken as a model for that data as a whole. Then, *with respect to this model*, one can work out how ‘likely’ more specific *selections* from that data would have been. For example, a very common situation in all kinds of empirical work, including multimodal investigations, is to explore whether two sets of data are significantly different from one another – i.e., if some kind of manipulation or some particular properties of the objects or performances about which data was gathered made a reliable difference; we illustrate this in the case studies below. To test this, we first make the assumption that there is *no* difference to be found because, if this were indeed the case, then a single description, or model, should apply to both sets of data. This assumption of no difference is called the *null hypothesis*.

Note that semiotically (and statistically), making the assumption that there is only one model is the more ‘conservative’ option – in general, one should aim for the least number of descriptions necessary to cover some body of data. Next, given this assumption, a statistical test calculates a specific value on the basis of the data that was actually found; how these specific values are calculated will be described below. Then, finally, this calculated value is placed against the particular probability curve selected for the data, just as was done with our value ‘9’ above with respect to any of the curves in Figure 1. If the value calculated is likely ‘by chance’ anyway (e.g., lies less than 2 standard deviations from the mean), then a single model, or description, cannot be ruled out and so one has probably not found a reliable difference; if, on the contrary, we find a configuration that was *not* likely according to the assumed model, then one is well on the way to discovering something that is reliably ‘there’ and not just a random event: i.e., we

have found something *distinctive* because it is unlikely that the situations characterised are subsumed by a single model. To predict outcomes reliably we would be better served by distinct models or descriptions; this is called ‘rejecting the null hypothesis’.

Much of the work in performing statistical tests to support argumentation thus comes down to the process of constructing models and seeing how likely it is that those models reliably reflect the data. We give examples of some very common models and the tasks for which they are best suited below. In addition, it will have become clear that there is a range of calculations to be performed on the basis of any data in order both to find models and to evaluate them. There is, however, no reason nowadays for these calculations to be done ‘by hand’. Powerful and freely available computer programs already have all the tests that one might want to perform implemented within them. In the next section, therefore, we describe how to make the most of the considerable resources now available. It is *strongly* recommended that one begins to make use of these tools as they turn rather tedious (and sometimes quite complex) calculations into easily accessible procedures.

3 Avoiding calculations: Using R

Although spreadsheet programs such as Microsoft’s Excel, Apple’s Numbers, and OpenOffice’s Calc all include a considerable body of ready-built calculations for statistics, we recommend here the use of a rather different software tool: a programming language called ‘R’ (R Core Team 2016). R is free, open source and includes a constantly growing collection of solutions contributed by researchers working at the state of the art in their fields. This means one is never dependent on ‘hidden’ code for performing any kinds of statistical procedures. Instructions for downloading and installing R are readily available on the web (search for: ‘installing R’); we also provide more details and all the examples discussed below on the accompanying webpage: <https://github.com/thiippal/stats-for-multimodality>.

Compared to spreadsheet programs, which are often ‘mouse’-based, use of R is rather different in that one types commands to perform the tests or calculations that one needs: we will show these in a moment and apply further such commands in specific examples below. There are several positive side-effects of this, despite the fact that being presented with menus in which options may be clicked may superficially appear simpler.¹ One side-effect is that one is much more likely to *learn what one is doing*. The situation here can be likened to always being shown the way to some destination by an automatic navigation system. Typically, use of the automatic system does not lead to one ‘knowing’ the way because one does not have to pay attention to landmarks and decision points – which can then leave one quite lost if something in the environment changes or one wants to go to a different location (even one that might be quite close to the previous one). Using mouse-based systems are, in fact, better suited to users who already have a good grasp of the underlying statistics because they will know what the presented alternatives mean. In contrast, when we use R, we put together the statistical tests that we want to perform ourselves; there is far less that is ‘hidden’. This becomes increasingly important as statistical tests become more complex (and more revealing), where it becomes natural to turn to explicitly constructing *descriptive models*, which is not something that is well supported by pre-programmed statistical tools.

A second positive side-effect of moving away from the mouse-based paradigm is that it means that *there is always an exact record* of precisely what tests were done on which data. This exact record can then be used by others for building on any results and checking that appropriate methods have been applied; in principle, this encourages not only publishing the results of data analyses but also the precise methods by which those results were achieved. This practice is becoming increasingly common and, indeed, increasingly expected in order to maintain transparency concerning claimed results. McElreath (2016) states this even more strongly, contrasting the approach taken in R with ‘graphical’ point-and-click spreadsheet programs thus:

“It is also a basic ethical requirement of science that our analyses be fully documented and repeatable. The integrity of peer review and the cumulative progress of research depend upon

¹Positions can differ on this issue of course; but, to be clear, we reject utterly the following position espoused by Eddington (2015: xvi) in his introduction to statistics for linguists using SPSS: “I understand that this excludes the program R. . . . I am aware that in comparison to SPSS, R is more powerful, produces better graphics, and is free. However, its command line interface makes the learning curve much steeper in comparison to programs with graphical user interfaces, and it is also quite intimidating for many computer users who are accustomed to point and click programs. Given the introductory nature of the present book, SPSS is a much gentler way of introducing people to statistics.” We will leave the reader to decide whether the command line interface is so complicated; what is not a matter of personal taste, however, is the fact that learning to click various buttons in SPSS will *not* help you understand much about statistics – quite the contrary: too much is presented too soon. Using SPSS then becomes more a question of correctly ignoring buttons that one has not yet understood the sense of!

it. A command line statistical program makes this documentation natural. A point-and-click interface does not. Be ethical.” (McElreath 2016: xiv–xv)

In this spirit, then, we show both here and on the accompanying webpage how all of the statistical tests and information we discuss can be produced in the fully transparent style that R supports and encourages.

The kinds of R commands that we use in this article will, however, remain exceedingly simple. Essentially they are of a single kind: calculations required for some statistical procedure are generally given a mnemonic name and so all that is required to use such a procedure is: (i) to type its name and (ii) to say which data the procedure should be applied to. Data sets are generally given labels so that we can easily refer to them. As our first example, therefore, we can take the task of finding the average of some set of values which we have previously made available to R and associated with the name `data`; on the accompanying webpage we show how it is straightforward, for example, to export data from spreadsheet programs so that they can be used directly in R; the only requirement is that the data is organised cleanly – we discuss what this means below. Calculating the average is then achieved simply by typing the following:

```
mean( data )
```

The name before the opening bracket is the procedure to be performed (in this case ‘mean’ for average) and the name within the brackets is the data the procedure is to be applied to. It makes no difference what this data is and how much of it there is; to make the example concrete, we took a set of 2193 measurements of the degree to which a speaker in a film lowered their jaw during a particular performance obtained automatically from facial expression analysis software and associated this data with the label `data`. R then prints the result:

```
[1] 6.925624
```

The number in square brackets just says that there was one number in the result; as we will see in examples below, R is most often used to produce results for entire sets of data, so we might have more than one result – but that is not the case here. The average of our 2193 data points was 6.925624.

Going further, calculating the standard deviation to see how far the data spreads out as described in the previous section is no more complicated. The name given to the standard deviation calculation in R is `sd`, so we can do the following:

```
> sd( data )  
[1] 17.50925
```

From now on we use the convention that lines preceded by `>` are the lines that you need to type, and the other lines are how R responds. So here we can see that this data set has quite a wide spread, as its standard deviation is large compared with the mean. This should already lead us to think more critically about what is being done with the data: does it make sense, for example, to say what the *average* amount of the jaw being dropped is? Sometimes the jaw is dropped considerably, when something is being said, sometimes it is not, when nothing is being said: so simply taking an average might not be sensible. This shows how it is always important to think carefully about what one is doing with data and never to blindly apply procedures. The procedures will always calculate a result, but deciding whether or not that result is sensible is where analytic control must remain with the researcher.

The last aspect of R that we need to introduce before proceeding is that, as mentioned above, a very large number of ready-made solutions and methods for calculating and visualising statistics results are written in R by researchers and then made available to the entire research community. These solutions are kept in what are called ‘libraries’ or ‘packages’. In order to use any of these solutions, we have to tell R that they are to be downloaded and made available for use in your own R programming environment. For example, one package we will use below allows us to make statements about how much data is required to address a particular question to get statistically significant results: this is the package called *pwr* standing for ‘power’ (Champely 2018). To use this package, one first has to obtain it, which is done by telling R:

```
> install.packages( "pwr" )
```

This only needs to be done once; subsequently, one can use all of the solutions contained in the package simply by writing:

```
> library( pwr )
```

Again, this is an important contribution to documentation, as it makes explicit precisely what tools one is using in order to derive some results; we use several such packages below. In the next section, we use R to apply the general statistical properties introduced in the previous section to several multimodal and visual studies. This will show how the application of statistical methods can support research as well as suggesting that this is nowadays far from complex.

4 Case studies

Multimodality researchers often wish to compare communicative situations or artefacts. Whether comparisons are drawn across cultures, over time or between artefacts or situations, they can be supported using statistical methods, as long as the framework provides analytical categories that can be reliably applied to a representative body of data, and the *results can be counted*. A count can tell us if a particular feature occurs more or less frequently, but only statistical tests can tell if there is actually a difference between categories. We begin with two examples of finding differences in two rather different kinds of data: both very common in multimodality research. We also address several common questions that always arise when considering statistical analyses, such as the kinds of tests to adopt and the issue of just how much data is necessary.

4.1 The use of text in infographics

In our first example, we consider a situation drawn from some current research on diagrams and infographics, in which blocks of texts have been found to occur in rather different ‘visual’ contexts. In network diagrams and schematic diagrams, for example, there appear to be differences in the amount of written language used. We might then hypothesize a trade-off between these expressive forms, with network diagrams needing to rely on written language more frequently than their highly conventionalized schematic counterparts. Instead of arguing this with examples as might commonly be done, we show here how to ascertain whether this hypothesis finds *statistical* support.

We consider a data set including 131 network diagrams and 21 schematic diagrams; for each diagram, we record the number of text blocks used as a kind of ‘measurement’. We could therefore apply the same approach whenever we have two situations to compare and some dependent property that we can give a value to. For example, we might want to know if different types of diagrams are read more or less quickly, in which case our measurement would be reading times. Or we might want to know whether people look at the text blocks more, in which case our measurements might be ‘gaze durations’ measured during an eyetracking experiment. All of these would lead to a similar statistical situation: i.e., we have two sets of measurements, one for each condition, and we want to know if the conditions make a difference: i.e., if the behaviour in one of the conditions is significantly different to the behaviour in the other condition. As suggested above, one intuitive way of characterising the kind of statistical test that we want to perform here is to see how likely it would be for the two sets of measures being compared to be compatible with a single model. When working with two sets of measurements, as in the current example, there are several tests that one might apply, depending on just how the data examined ends up being ‘shaped’, or distributed. The first step is then to look at the data to see how the values in the data are distributed; on the basis of this, it is possible to choose an appropriate test, that is, a test whose ‘model’ has a shape that is compatible with the shape found in the data.

The essential criterion to apply here is the extent to which the distribution of the data approximates the kinds of normal distributions that we saw above. If the shape of the data is like this (and, as we noted above, many measuring processes will produce such a curve), the standard test to apply is called the t-test. The t-test consists of a particular formula that calculates the ratio between the difference between two groups as compared with the differences within those two groups. The larger the ratio, the more different the groups are; the lower the ratio, the more similar. The value of this ratio is called the t-statistic and this is used to determine just how likely it would be for both sets of data to be drawn from a single model in precisely the same way as was described above using the normal distribution (cf. Figure 1). The t-test will produce reliable results when the two groups of data are normally distributed, the samples have around the same size, *and* similar standard deviations – if these conditions do not hold, then the formula just described is stretched too far.

We can easily check whether these conditions hold for our current data as follows. First, the sizes of the two samples are quite different (with 131 network diagrams and only 21 schematics), so we will certainly need to be wary of the t-test in any case. Second, we saw how to calculate standard deviations above, so we repeat this here for our two data sets. If we assume that we have the raw figures for network

diagrams in a table called ‘networks’, and the corresponding figures for the schematic diagrams in a table called ‘schematics’ – i.e., we have two tables simply listing the numbers of text blocks found in each individual diagram, then:

```
> sd( networks ); sd( schematics )
[1] 5.026728
[1] 2.943111
```

tells us that the shape of the curves of the two data sets are also quite different: comparing with Figure 1 above, the networks data are far ‘flatter’ and more spread out (SD=5.03) than the schematic diagrams (SD=2.94). Finally, R also offers several straightforward ways to check whether the data is (sufficiently) normally distributed; there are both visual inspection methods and statistical tests for this as well. The statistical tests of normality can be very picky, however – i.e., readily rejecting distributions that are ‘almost’ normal – but they are easy to apply. The ‘Shapiro-Wilk Normality’ test, for example, gives us the probability that we would be wrong to assume that the data tested is *not* normally distributed. We can perform this test for our network diagram data works as follows:

```
> shapiro.test( networks )

Shapiro-Wilk normality test
data:  networks
W = 0.97515, p-value = 0.01661
```

The probability (given by the ‘p-value’) is quite low – which means that assuming that the data is not normally distributed might not be a bad decision. we show a more straightforward visual method that leads us to the same decision on the accompanying website. In the present case, therefore, taking all of these results together means that the t-test is not a suitable choice for our data.

Whenever we reject the t-test on these grounds, an alternative for checking whether there is a difference between two sets of measurement data is called the ‘Wilcoxon two-sample’, or Mann-Whitney, test. Performing it in R is again straightforward. We write:

```
> wilcox.test( networks, schematics )

Wilcoxon rank sum test with continuity correction
data:  networks and schematics
W = 2285, p-value = 1.143e-06
alternative hypothesis: true location shift is not equal to 0
```

This applies the formula hidden behind the name ‘wilcox.test’ to the two data sets and returns the values we need in order to come to a conclusion concerning whether a single model suffices (i.e., the null hypothesis is not rejected and there is no significant difference between the conditions) or whether the assumption of a single model is so unlikely as to be rejected. Since the formula here is slightly more complicated than that described informally above for the t-test, it is good that R hides that information, unless we want to see it, making application of this test just as straightforward as applying any other test.

The first number we see (labelled by ‘W’) is a value returned by the formula to work out the probability that we would be wrong to reject description by a single model. This probability is, again, given by the ‘p-value’. The number given for the p-value is R’s way of writing 1.143×10^{-6} , i.e., 0.000001143. This is a probability so small that it can be ruled out as a serious contender. Often when reporting statistical results, one seeks to be ‘sure’ at either 95% or 99% levels: i.e., achieving probability, or ‘p-values’, of 0.05 and 0.01 respectively. The more strict meaning usually given to ‘p-values’ is then that they characterise the probability of falsely rejecting the null hypothesis: i.e., of *wrongly* taking the step of assuming that we cannot remain with a single model; this was also how we read the result of the Shapiro-Wilk test for normality above, where the value was below 0.05 but above 0.01. Naturally, to reject a null hypotheses, we want this probability to be as low as reasonably achievable so as to be sure as possible that we are not making a wrong decision. In the case of our diagram data, therefore, the result is said to be highly statistically significant as the p-value is far below the standard levels: there is virtually no chance that the two kinds of diagrams are behaving similarly with respect to their use of text blocks. And, drawing on the means and standard deviations again, we see that the average use of text blocks in network diagrams is far higher (now we can say statistically significantly higher) than the average use of text blocks in schematic diagrams.

If, on the other hand, our data had passed the criteria given above for using the t-test, the R code that we would need only has to change the name of the test. So we would write:

```
> t.test( networks, schematics )
```

This prints a similar report to that of the Wilcoxon test, giving the formula's value (in this case called the t-statistic) and the p-value calculated on that basis. Statistical tests, therefore, normally work in this way, producing values (e.g., W or the t-statistic) that can be placed against the corresponding models, or curves, for those tests in order to produce p-values. If the value produced by a test is high enough, we know from the corresponding curve that this would be extremely unlikely to occur by chance and so we can safely reject the assumption of a single model: i.e., the null hypothesis that there is no difference. Nevertheless, as we have now discussed, all statistical tests have boundary conditions, outside of which calculated results may not be relied upon. R would not stop us applying the t-test to our data here, for example, and would dutifully report values: it is always up to the researcher to consider whether this is a sensible thing to do, which is what the criteria given above let us decide. In more traditional introductions to statistics, the sheer variety of boundary conditions that might need to be considered for each test can appear very much as 'arcane knowledge', which just needs to be accepted as given. Now, we can place a rather different slant on this situation: all the boundary conditions do is describe the range of models that the test has been found to work well for. Clearly, therefore, if we can already see that our data has properties that contradict, or poorly align, with the model of any particular test, then that test is going to perform poorly. Again, a semiotic view of attempting to find the most accurate descriptions of any sets of data offers a good foundation for thinking through any experimental or observational situation for which we want to perform analysis.

Finally, just as with citing literature in research articles, there are established conventions for reporting the results of statistical tests as well; these should be adhered to when publishing results. Reports following these conventions tell just what test was done on how much data, and what the principal results of the test were. For our Wilcoxon/Mann-Whitney test on the diagram data, one would write something of the form:

In a sample of diagrams consisting of 131 network diagrams and 21 schematic diagrams, a Wilcoxon-Mann-Whitney test showed that significantly more text blocks were used in network diagrams ($M = 9.8, SD = 5.03$) than in schematic diagrams ($M = 4.48, SD = 2.94$), $W = 2285, p < 0.000$.

The p-value is reported as simply being very low because it is far below standard 0.05 and 0.01 levels.

Following these steps, which have taken far longer to read and to report than they did to actually perform, we can proceed to more specific, or more explanatory, investigations of this difference in usage: all the statistical test has told us is that there is a difference and so further research might be beneficial. If we had not found the difference to be significant, we would be in a far less secure position concerning our hypothesis. Nevertheless, failing to find significance is *not* an indicator that there is no difference – only that our results have not demonstrated this yet and so caution might be advised. We would certainly not be in a position to report that a difference has been observed, because it was not. It is very worthwhile, therefore, getting into the habit of performing such tests whenever appropriate and of thinking how to obtain data for which tests could be applied when designing any analysis at all. In short, all kinds of measurements will exhibit variation, but statistical tests can reveal whether this variation is simply random noise or systematic variation, which might then potentially be meaningful.

4.2 Differences in the semiotic resources used in Chinese and UK newspapers

The next example uses a slightly different kind of data and revisits a previous study by Kong (2013), who presents a corpus-based study comparing 55 pairs of news items from two tabloids, *The Oriental Daily* (China) and *The Sun* (UK). The news items consisted of 4178 analysis units, each of which was annotated for its multimodal characteristics. Kong (2013: 183) states that “there is a significant difference in the overall distribution of semiotic resources in the two newspapers.” The phrase ‘significant difference’ suggests statistical significance, which results in a delicate situation because this claim is not backed with evidence from statistical tests.

Table 1 sets out the figures from Kong's study. Kong gives these in a pie chart which is visually suggestive of a difference but is not sufficient on its own to claim significance (Kong 2013: 185). The statistical question to be addressed is whether there is a reliable difference in usage across the newspapers and across the categories of semiotic resources. As explained above, the ‘null’ hypothesis to be investigated

Table 1: The observed counts of semiotic resources used in two tabloids; data taken from Kong (2013: 185, Figure 5).

	<i>The Oriental Daily</i>	<i>The Sun</i>
Text	1551	1653
Photo-pictorial	560	255
Diagram	133	26
Totals	2244	1934

is that the counts show *no* difference: if we then find that the probability of the counts coming out in this way is extremely low (i.e., lower than 0.01 or 0.05), then we can confidently reject the hypothesis of there being no difference. The null hypothesis is always formulated in this way so that one can be as sure as (statistically) possible that one does not assume a difference mistakenly.

To begin, we return to Kong’s initial data and the values he reports. When gathering data, counts and measurements, it is always advisable to place such collections of results into *tables* as we have done in Figure 1. There are, however, two very different kinds of tables and it is important methodologically to be very clear on the functions they serve. When we are still collecting and collating data, we use *data tables*. These are used primarily for exporting results to any tools for exploring the data, performing statistical tests for differences, or for producing graphs. When we move on to presenting results for readers, we use *display tables*, which are solely for *reporting* or *visualising* research results. These are completely different functions and so need to satisfy different criteria. Data tables should respect a few basic principles. First, each entry in the data table, that is, each ‘cell’, should *only contain one type of information*. Aggregating information, such as counts and percentages into a single cell, e.g. “27 (46%)”, is not a good idea because then one cannot use the table for doing any further work, such as adding up the numbers in a column to see the totals, finding maximum or minimum counts, or even calculating further percentages. If one writes “27 (46%)” then any program just interprets this as a string of characters: it will not then know which characters correspond to numbers and which correspond to other information. Instead, dividing the counts and percentages into their own cells, or columns, makes them easily accessible to other programs (such as R) for performing further calculations.

The criteria for display tables, which are typically constructed by selecting specific scores from data tables, are that they should be designed appropriately for actual readers to get at the information they contain; hence they need to be compositionally appropriate, i.e. visually well-designed, to express their contents, grouping related information together in cells, using coloured backgrounds of various kinds, and so on. The temptation to place many pieces of information together in individual cells (sometimes to the point of confusion) often arises due to mixing up these two kinds of tables. Finally, note also that somewhere in any display tables used it should be possible to see the *raw figures* used in the statistical calculations: never just give percentages or other indications of proportions between categories. Knowing that something occurred two thirds of the time is a much more valuable fact if that happens 666 times out of a thousand than it is if it happens 6 times out of 10. Just giving proportions disguises this and makes it difficult to see the basis on which any tests reported are operating. Kong fortunately does provide these raw figures, making it possible to proceed to statistical evaluation.

Statistically, however, the situation under investigation here is completely different to the previous example of text blocks in diagrams. We do not have measurements of some single variable: instead we have counts of a variety of different kinds of data – i.e., the three kinds of semiotic resources investigated. This situation is extremely common in multimodality research. Categories are set up indicating some properties of some objects of analysis and then are compared concerning how often they occur across different kinds of data. Counts of categories of this kind constitute what is termed *categorical data* and one appropriate statistical test for this is known as chi-square (χ^2). We use chi-square whenever we wish to compare qualitative categories that have no intrinsic ‘ordering’ of their own: for example, ‘male’/‘female’, ‘given’/‘new’, ‘elaboration’/‘extension’/‘enhancement’, or ‘text’/‘photo-pictorial’/‘diagram’. When we count up how many cases of particular categories we have found, this is rather different to taking a measurement; whereas we might be able to ‘add’ counts (e.g., the total types of expressive forms), forming averages or means often makes little sense. Since the numbers in our data have rather different properties to measurements, therefore, the underlying formula for calculating chi-square is different to those of the tests we have used above. However, the procedure for the use of chi-square remains the

same as we have seen above: a formula specific to the test is applied and this returns a value, which in this case is called the chi-statistic. This value is then placed against the corresponding probability distribution for chi-square to obtain a probability indicative of the observed values occurring by chance. If that probability is sufficiently low, then it is unlikely that chance was involved and we can safely assume that there is a difference between the use of the categories.

As with the tests shown above, there is no need to do any calculations ‘by hand’. There are many computational resources available for performing these tasks, calculating all the necessary intermediate values and subtotals, in order to ascertain the final p-value. There are even several websites that perform these operations online. In all cases, all that must be provided is a clean data table containing the observations to be tested. If we use R to do the calculations, the first step is again to make the data known to the R environment. Because in our examples in the previous subsection there were many values, the tables were quite large and so would normally be read from a file. In the current case, however, Table 1 is small, i.e., just 2×3 , so we can just type in the values directly. The following consequently stores the values of the table in an appropriately shaped table called `kong_table`:

```
> kong_table = matrix( c( 1551, 560, 133, 1653, 255, 26 ), ncol = 2 )
```

This is read as follows: first, form a list of numbers beginning with 1551 and ending with 26 by using R’s ‘combine’ function, `c`; then shape this list into a `matrix` with 2 columns (given by the number of columns parameter: `ncol = 2`). This just consumes the numbers given to the matrix two-by-two, yielding 3 rows, and associates the resulting table with the name `kong_table`; this is also shown step-by-step on the accompanying webpage. Subsequently, to perform the chi-square test, all we need is:

```
> chisq.test( kong_table )
```

to which R responds:

```
Pearson's Chi-squared test
```

```
data: kong_data
```

```
X-squared = 167.31, df = 2, p-value < 2.2e-16
```

A new piece of information appears here: ‘df’, which is the ‘degrees of freedom’ involved. For tables such as that shown in Figure 1, this is obtained simply by counting the number of columns (2) and the number of rows (3) of the table, subtracting 1 from each count, and multiplying the counts together; the idea behind this is to characterise how many independently varying groups of data are involved. Thus we get $(2 - 1) \times (3 - 1) = 2$ degrees of freedom. The final “2.2e-16” is R’s way of writing 2.2×10^{-16} , i.e., a very small number indeed and so clearly much less than the usual targets of 0.01 or even 0.001.

One can therefore, indeed, safely reject the hypothesis that there is no difference across newspapers and semiotic resources, just as Kong suggests. The standard phrasing for this result, combining the name of the test, the degrees of freedom (2), the number of observations (4178), the value of the chi-square statistic and the final probability, is then:

A chi-square test of independence was performed on the semiotic resource types ‘text’, ‘photo-pictorial’, and ‘diagram’ across the two newspapers. The relation between these variables was significant, $\chi^2(2, N = 4178) = 167.31, p < .000$, showing an interdependence between newspaper and type of semiotic resource employed.

This tells us that there is a difference observable in the data, but does not yet pin down just which variables contributed to that difference. From the counts, we might, for example, assume that the difference lies more with the ‘photo-pictorial’ and ‘diagram’ distributions than with the amount of ‘text’. This can also be calculated statistically by examining the chi-square statistic for *each cell* in the data table: these are called *residuals*. Predictably, the highest of these contribute most to the overall dependence that has been found. R makes it easy to inspect these values and, when we do so, we find that all of the variables actually make strong contributions. So there are many significant differences and dependencies to follow up on in this study.²

²Note that, again, just as with the t-test above, we need to be careful that our data does not violate any of the assumptions that are built into the chi-square test. For linguistic corpus work perhaps the most significant assumption to be aware of that is often not given enough consideration is that the conditions need to be mutually exclusive and *independent* of one another. This is important because, as Kilgariff (2005) famously notes, in text corpora this is very rarely the case: for example, the selection of words in a text are *not*, in general, independent of one another! The extent to which this influences multimodal corpus studies will need careful monitoring and is a further reason to move towards more explicit modeling as we suggest throughout this paper. Gries (2005) also notes the importance both of providing *effect size* values, which we discuss below, and of using appropriate corrections when applying tests multiple times to the ‘same’ data.

4.3 How much data is enough?

When considering performing statistical analyses of the kinds we have now illustrated with the Wilcoxon-Mann-Whitney test, the t-test and the chi-square test, the question usually arises as to just how much data is necessary for carrying out analyses reliably. It is not uncommon, particularly in multimodality research, for researchers to present this as a reason (or excuse!) for not performing or reporting any statistical tests: i.e., it is suggested that there was not enough data. If, the argument goes, they are only examining a small sample, there is no point ‘doing statistics’. Although it can well be the case that one actually does have too few data to perform statistical tests reliably, this should always be approached as an open question that needs to be raised rather than simply assumed. In fact, in some cases the quantity of data necessary to be able to make some reliable statistical statements may be very much smaller than might have been thought. In such cases, moreover, it is always worthwhile considering this question *beforehand*, so that one does not, simply out of a lack of forethought, find oneself in the position of having a relatively strong result but just a handful of cases too few to report on statistical tests. Had one in such cases perhaps done just a few more analyses, or collected just a few more questionnaires, or evaluations, or whatever, that might have been enough for a reliable result of publishable quality.

The quantity of data required to obtain reliable results depends on one variable, and that is the so-called *strength* of the effect that one is studying. A strong effect will make its presence reliably known with a relatively small quantity of data; a weak effect will require considerably more data. It is then worthwhile thinking carefully about what it would take to set up the study one is planning in a way that makes the effect one is looking most likely to be a strong one, again pointing to the value of considering experimental design carefully at an early stage in any inquiry. For each type of statistical test there are ways of calculating the strength of any effect that they reveal. We can show this easily by returning to Kong’s data and the chi-square test that was performed there. We saw that the result was significant but how strong was the effect: i.e., was the effect significant but rather weak, or perhaps quite strong? For chi-square one of the standard measures of effect strength is called Cramer’s V. And, again, we do not need to know the formula involved in this but can go straight to the result using R (in this case, the necessary function is in Daniel Navarro’s ‘learning statistics with R’ package: ‘lsr’; we assume that this has been installed and loaded beforehand as described above):

```
> cramersV( kong_table )  
[1] 0.2001162
```

This tells us that although the result was significant, the effect size was rather weak. Values for Cramer’s V from 0.2 to 0.6 are considered moderate, while any value above 0.6 indicates a strong effect. Similar functions are available for other statistical tests as well; for the t-test, for example, the *lsr* package provides the function `CohensD`, which is used in a very similar fashion.

The consequences of a weak effect are that more data are required to show the effect at work with any reliability. This is useful because we can use it to determine just how much data is necessary. To show this we employ what is called *statistical power*. Power analysis measures how likely we are to detect an effect if an effect is there and thus correctly reject the null hypothesis of there being no difference. A statistical test with a power of 0.8 translates to an 80% chance of detecting an effect if there is one, and this is a value commonly used as a desirable target. If, on the other hand, an experiment or measurement has low power, then we might just not have enough data to rely on our results, even if the tests performed suggest significance. How statistical power is calculated again depends on the specific statistical test used. For a chi-square test, measuring statistical power requires knowing four out of five variables – i.e., if we know any four of the following five variables, we can work out the fifth: the effect size, commonly labeled w ; the number of observations (N); the degrees of freedom (df); the significance level targeted (traditionally selected at 0.05 or 0.01); and the power required, i.e., the probability of detecting an effect that actually exists (traditionally set at 0.8).

Now, for the example of Kong’s data we already have these values and so can explore just how much data would be required to be 80% sure that a deviation from the null hypothesis would be detected at some selected significance level, if such a deviation exists. To do this, we use the ‘power’ package *pwr* of Champely (2018) mentioned above. This package makes available several of the standard power analyses proposed in a classic work in the field, that of Cohen (1988). All we need to do is provide the values of the parameters that we consider fixed and the package will calculate the remaining value. This is achieved for chi-square with the following single line of R code (again assuming that the ‘pwr’ package has been installed and loaded). The value we want to find, in this case the number of observations necessary, is indicated by giving ‘NULL’ as the value:

```
> pwr.chisq.test( w = 0.20012, N = NULL, df = 2, sig.level = 0.01, power = 0.8 )
```

Chi squared power calculation

```

      w = 0.20012
      N = 346.6015
      df = 2
sig.level = 0.01
power = 0.8

```

This tells us that to achieve a result at the (quite severe) 1% significance level (0.01) we would need to have at least 347 samples (the value of N reported) to be 80% sure of detecting a result if there is one. This is rather fewer than the 4178 observations that Kong actually had and so we can continue to be reasonably confident that the effect found in that data was reliable: we would, in fact, have found this effect even with considerably less data.

Knowing the likely effect size for some phenomenon is consequently a very valuable piece of information. And for this reason, it is increasingly the case that effect sizes are expected when publishing results. This makes it possible for other researchers to consult the literature and to come up with more motivated estimates of the amount of data they will need in order to show statistically significant results if there are any to be found. If a phenomenon is being addressed for which there are no existing effect size results in the literature, then publishing this information in any new study is both advisable and useful for the broader research community. In summary, then, although it is commonly thought that the more certain we want to be about the reliability of our results, the more data we need, just how much data can be calculated on the basis of effect size. If we investigate a phenomenon with a strong effect size, that number may turn out not to be so large after all. For example, applying the R code above for some situation that has been calculated previously to exhibit an effect size of 0.6 (commonly assumed to indicate a strong effect) gives back a sample size N of only 39. This means that to be 80% sure of finding a result if there is one, we would only need 39 items – which is a number already lying within the capabilities of many BA projects.

4.4 Is the description of data reliable or not?

So far we have applied statistical tests to data that has already been classified and quantified in some way. There are equally important questions, however, concerning the process of classification itself. Whenever we use some abstraction to classify forms of multimodal communication, such as an annotation schema or other set of descriptive categories, the question always arises as to how *reliable* or consistent that description is. Whenever annotations of data are reported in articles, reviewers nowadays will increasingly as a matter of course ask what steps were taken to ensure that the annotation on which the analysis is based is itself reliable. If a description of data used in analysis is itself questionable, this can waste a considerable amount of time and effort as well as lead to conclusions that turn out to be untenable. As Artstein and Poesio put it:

“Researchers who wish to use hand-coded data – that is, data in which **items** are labeled with **categories**, whether to support an empirical claim or to develop and test a computational model – need to show that such data are **reliable**.” (Artstein and Poesio 2008: 556)

Artstein and Poesio are primarily concerned with linguistic analysis, but similar questions can be raised for *any* types of coding schemes, of which there is a huge range of possibilities (cf. Saldaña 2016).

In general, reliability is evaluated statistically by measuring what is variously called *inter-annotator agreement*, *inter-coder consistency*, or *inter-rater reliability*; we use these terms interchangeably here. All such tests build on the assumption that annotated data can be considered reliable if annotators *consistently agree* on how an annotation schema should be applied to a particular body of data (Artstein and Poesio 2008: 557). Consistent agreement gives a good indication that independent annotators will produce the same kinds of descriptions and so these can stand as a suitable foundation for further evaluation and research. Although the simplest measure of agreement might appear to be *percentage agreement* – and, indeed, one still sees this being reported in published papers – this is problematic in many respects. Consider the following situation: imagine that one of a pair of annotators does not take the task seriously at all and always makes the *same* choice: we can still calculate a percentage agreement but this then evidently has nothing whatsoever to do with the annotation scheme; it simply reflects the proportion of times that the other annotator happened to pick the same label as the label that the first

annotator always picked. Proper inter-annotator agreement tests go further by attempting to rule out the possibility that the annotators may happen to agree by chance. To do this, they look not only at how much agreement there was but also at how much of the annotation scheme was actually used. If a broad range of possibilities is being used and the annotators agreed about these, then it is unlikely that this would be due to chance.

Several standard measurements for reliability avoiding chance agreement are available and are already broadly cited in linguistic publications. Here we will illustrate the use of one of the most common of these tests, Cohen’s kappa (or κ) statistic (Cohen 1960); different tests attempt to deal with a range of further issues, such as missing data, or comparing more than two annotators, but the basic method of preparing data remains the same as we describe here. As an example, we show the reliability calculation drawing on Hiippala and Orekhova (2018), who describe an annotation situation in which two annotators had to assign pre-defined discourse relation labels to a sample set of 545 diagram items drawn from primary school science. Annotation tasks of this kind are common in qualitative multimodal research and so the method described here is of very general applicability.

In order to evaluate the reliability of an annotation scheme, it is crucial that the task of annotation is carried out independently by all coders: that is, the annotators involved must make their own choices of annotation labels, without discussion or changes depending on the decisions of the others. Evaluating reliability therefore typically follows a preceding training stage, where discussion will be both necessary and useful. Only when the annotators are deemed sufficiently proficient is it then appropriate to conduct a reliability check. And, following the reliability check, if an insufficiently high score is achieved, there would need to be more training or refinements to the annotation scheme *before* proceeding to apply the scheme to real data. When both annotators have annotated the same sample of test data, their results can be organised in a two-column table with the labels selected for each item shown side by side. For the discourse labeling task, this might look as shown in Table 2. This is all the information that is necessary for calculating the reliability of the coding.

Table 2: Example of first four lines from a set of two independent codings of some data. Each annotator has their own column, and the rows correspond to the individual items analysed in the data. If the rows contain the same labels, the annotators agreed; if not, they disagreed.

	<i>Annotator 1</i>	<i>Annotator 2</i>
<i>item 1:</i>	EFFECT	EFFECT
<i>item 2:</i>	EFFECT	EFFECT
<i>item 3:</i>	PROPERTY-ASCRPTION	PROPERTY-ASCRPTION
<i>item 4:</i>	ELABORATION	IDENTIFICATION
$\vdots$	$\vdots$	$\vdots$

Calculating many of the reliability measures is relatively straightforward by hand, but rather tedious when the number of items to be checked is larger. Again, using R makes this easy. The table of results has to be organised into a tabular, machine-readable form – a data table – and loaded into the R environment as usual. For our example we will assume that we have the 545 pairs of annotations loaded into an R data table called ‘data’. Then, another R library called *irr* for ‘inter-rater reliability’ (Gamer et al. 2019) provides multiple measures of agreement, including Cohen’s kappa, that can be directly applied to the data table. For kappa, we simply use the library’s function `kappa2` to get the following results:

```
> kappa2( data )
Cohen’s Kappa for 2 Raters (Weights: unweighted)

Subjects = 545
Raters = 2
Kappa = 0.725

z = 36
p-value = 0
```

As usual, the function includes in its report information about the data it is using so that we can check (and document) what we are doing as well as some further statistical confidence measures about the results produced. What we are most interested in here, however, is simply the kappa score of

0.725. The theoretical values for Cohen’s kappa range between +1 for perfect agreement and −1 for perfect disagreement. Values above 0.6 are often considered necessary to be at all acceptable, with values above 0.8 indicative of excellent agreement. The result with the present data therefore suggests that the proposed categories can indeed be reliably applied to describe the discourse relations in diagrams. When publishing results based on such annotations, one should consequently also give kappa scores as evidence that the data analysis was reliable. As with all statistical methods, however, there are limitations here which may also need to be considered (see e.g. Di Eugenio and Glass 2004), but conducting such reliability tests is always an important step.

5 Modelling data and constructing theories: the Bayesian view

So far we have primarily presented statistical tests as ways of determining to what degree we can be sure that rejecting a particular model, typically the model assuming no difference that constitutes a null hypothesis, is a justified move. We also mentioned above, however, that null-hypothesis significant testing was just one kind of statistical inference. In essence, this kind of statistical reasoning takes the ‘long view’ on probabilities: that is, obtaining a particular p-value from a test means that if we were to carry out many tests over a long period, then the selected cut-off points for significance that we measure p-values against are the probabilities of falsely rejecting null hypotheses. We keep these low so that, in the long-term, we will not have rejected null hypotheses wrongly very often. If we look at our hypotheses as theories, then doing this kind of testing is working out *the probability of the data* given that the assumed theory is true: so a p-value of 0.03 means that our observed data should occur 3 or less times out of a hundred if our assumed null hypothesis were indeed the case. Many statisticians have argued that selecting arbitrary cut-off points for significance in this way is odd (cf. Field 2016: 364–370): after all, why should a test yielding a p-value of 0.051 be thrown out as not significant, whereas a test yielding a value of 0.049 is considered significant (at the 95% level)? The answer is precisely the long-term view: if we allow higher values, then mistakes are going to add up.

The other approach to inferential statistics we mentioned was Bayesian and this takes a rather different perspective. Bayesian inferential statistics operates by progressively incorporating ‘prior knowledge’ about a situation of interest and considering data as evidence to be weighed with respect to that prior knowledge. Data thus leads to a new state of knowledge, which can in turn be confronted with further evidence as more data is considered, and so on. In essence, then, probabilities in the Bayesian view are probabilities of a theory, or model, being true given the observed data. Doing statistical inference in this paradigm generally starts with an under-specified, or uncertain, theory or model and progressively uses data as a way of reducing uncertainty about that model. Since, as suggested at the outset, we are often most interested in working out theories or models of what is going on, the Bayesian view is rather more natural: it is the data that we know, not the theory, and so refining a theory on the basis of data would appear to be the more logical procedure.

This procedure is, however, computationally far more expensive than the simpler fixed cut-off point techniques of null-hypothesis significance testing and so has not been so broadly applied. But with increasingly ready access to computational techniques, this situation has changed and now, for example using R (cf. McElreath 2016), performing Bayesian statistics is little more complex than more traditional significance testing. Space precludes us going into any more details, but hopefully this article will have whetted your appetite sufficiently to explore further yourself.

6 Conclusion: a set of recommendations

In this article, we have suggested that in many respects the current situation in multimodality research resembles that of linguistics in the 1980s and early 1990s when genuinely empirical work was largely limited to sociolinguistic and corpus-based research, two areas where the complexity and variability of the data has always demanded statistical methods, and to psycholinguistics, inheriting research paradigms from experimental psychology. Since then, the value of performing basic statistical tests for empirical studies has been recognized for a broad range of linguistic questions and has become relatively common. As Carletta characterized the situation in the mid-1990s:

“At one time, it was considered sufficient when working with such judgments to show examples based on the authors’ interpretation. . . Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand, and reliably make, the judgments

underlying the research. This is a reasonable requirement, because if researchers cannot even show that people can agree about the judgments on which their research is based, then there is no chance of replicating the research results.” (Carletta 1996: 249)

Such requirements, and the means for meeting them, are now part and parcel of the empirical linguist’s toolset; they need now to be equally accepted in research on multimodality and visual communication as well.

To help take this step, we have attempted in the current article to make some of the most common statistical techniques necessary in multimodality and visual research not only more approachable, but also more intelligible by relating them to more semiotic notions of models and descriptions. Very generally, good statistics relies both on descriptions being appropriate for the data at hand and on their possessing certain properties supportive of drawing inferences. It is these properties that make the descriptions both reliable and powerful. In a very real sense, then, these descriptions function as Peircean *metaphors* (Peirce’s third variety of iconic signs), whereby properties of one description may be used to infer properties of another. Modern approaches to statistics increasingly take this ‘model building’ perspective when working with data and so this can be made to dovetail very well with a semiotic view. Much modern statistics is consequently concerned far less with testing whether hypotheses are ‘correct’ or not (an impossible goal in any case) and more with constructing reliably predictive descriptions of data: also an inherently semiotic goal.

The article has nevertheless barely been able to scratch the surface of statistical methods. What we have shown, however, is that the current state of the art in computational tools for statistical analysis makes it straightforward to use most commonly applied statistical techniques. Learning the very basics of R or some other programming language is often sufficient, and will already pay considerable dividends. Many libraries are available, often performing multiple statistical tests at once as well as providing documentation of the techniques applied. For the researcher, crucial skills remain knowing how to prepare and explore data and to interpret results. Using computational tools requires, for example, that data be well organised and machine-readable. This should never, however, be seen as a pointless overhead: carefully organised research data is methodologically desirable in its own right – not only does it better support exploration and hypothesis building, but that data is also then ready to be released in connection with academic publications, allowing others to reproduce results, to learn from research designs, and to extend studies more readily to new or related sets of data.

In addition to finding reliable patterns, we have also shown that statistical methods in multimodality research offer ways of ascertaining to what extent any categories proposed are themselves reliable. It is telling that much work in visual communication and multimodality to date has not investigated the reliability of its own concepts. Such work cannot then, by its very nature, be used to make general claims about how data will actually pattern; for this the empirical testing necessary still remains to be done. It is therefore important for work of this kind to be reported realistically as *developing hypotheses* to be tested further rather than suggesting that results have already been achieved. Such hypotheses can be valuable in their own right, but still need to be clearly recognised as such. By applying the techniques described in this paper, we hope to make it easier for researchers to engage in these critical tasks of self-reflection as well.

Further reading

Whereas even 5 years ago the majority of ‘introductions’ to using statistics for any kind of linguistic or related studies would have been unapproachable for any without at least some mathematical, or even better, some statistical, training, this situation is now changing radically. There are several books which take a deliberately straightforward and practical approach, often directly supported by giving R code that even the complete newcomer can execute, examine the results, and so see statistical processing in action. Newcomers may previously have been daunted even by the word ‘statistics’ in the title, let alone reference to a programming language. We hope that this article and its supporting webpage can act as a bridge, demonstrating not only that the step is worthwhile but also that it is now no longer that large a step in any case. Books of this kind that we recommend include: Levshina (2015), Gries (2016) and Winter (2020); there is even an extensive introduction (partly) in the form of a graphic novel: Field (2016).

References

- Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2007), ‘The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena’, *Language Resources and Evaluation* **41**(3), 273–287.
URL: <http://dx.doi.org/10.1007/s10579-007-9061-5>
- Artstein, R. and Poesio, M. (2008), ‘Inter-coder agreement for computational linguistics’, *Computational Linguistics* **34**(4), 555–596.
- Baayen, R. H. (2008), *Analyzing linguistic data. A practical introduction to statistics using R*, Cambridge University Press, Cambridge.
- Bateman, J. A. (2018), ‘Peircean Semiotics and Multimodality: Towards a New Synthesis’, *Multimodal Communication* **7**(1), aop.
- Bateman, J. A. (2019), Afterword: Legitimizing Multimodality, in J. Wildfeuer, J. Pflaeging, J. Bateman, O. Seizov and C. Tseng, eds, ‘Multimodality: Disciplinary Thoughts and the Challenge of Diversity’, de Gruyter, Berlin, pp. 297–321.
- Bateman, J. A., Wildfeuer, J. and Hiippala, T. (2017), *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*, Mouton de Gruyter, Berlin.
- Caple, H., Bednarek, M. and Anthony, L. (2018), ‘Using Kaleidographic to visualize multimodal relations within and across texts’, *Visual Communication* **17**(4), 461–474.
- Carletta, J. C. (1996), ‘Assessing agreement on classification tasks: the kappa statistic’, *Computational Linguistics* **22**(2), 249–254.
- Champely, S. (2018), ‘pwr: Basic functions for power analysis’.
URL: <https://cran.r-project.org/package=pwr>
- Cohen, J. (1960), ‘A coefficient of agreement for nominal scales’, *Educational and Psychological Measurement* **20**(1), 37–46.
- Cohen, J. (1988), *Statistical power analysis for the behavioral sciences*, 2 edn, Erlbaum, Hillsdale, NJ.
- Di Eugenio, B. and Glass, M. (2004), ‘The kappa statistic: A second look’, *Computational Linguistics* **30**(1), 95–101.
- Eddington, D. (2015), *Statistics for Linguists: a Step-by-Step Guide for Novices*, Cambridge Scholars Publishing, Newcastle-upon-Tyne, UK.
- Field, A. (2016), *An Adventure in Statistics: the Reality Enigma*, Sage, London. Illustrated by James Iles.
- Gamer, M., Lemon, J., Fellows, I. and Singh, P. (2019), ‘irr: Various coefficients of interrater reliability and agreement’. R package version 0.84.1.
URL: <https://CRAN.R-project.org/package=irr>
- Gries, S. T. (2005), ‘Null-hypothesis significance testing of word frequencies: a follow-up on Kilgariff’, *Corpus Linguistics and Linguistic Theory* **1**(2), 277–294.
- Gries, S. T. (2016), *Quantitative Corpus Linguistics with R*, 2 edn, Routledge, London and New York.
- Hiippala, T. and Orekhova, S. (2018), Enhancing the AI2 Diagrams dataset using Rhetorical Structure Theory, in ‘Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)’, European Language Resources Association (ELRA), Paris, pp. 1925–1931.
- Jänicke, S., Franzini, G., Cheema, M. F. and Scheuermann, G. (2016), ‘Visual text analysis in digital humanities’, *Computer Graphics Forum* **36**(6), 226–250.
- Kilgariff, A. (2005), ‘Language is never, ever, ever, random’, *Corpus Linguistics and Linguistic Theory* **1**(2), 263–276.
- Kong, K. C. (2013), ‘A corpus-based study in comparing the multimodality of Chinese- and English-language newspapers’, *Visual Communication* **12**(2), 173–196.
- Kress, G. and van Leeuwen, T. (2006), *Reading Images: the grammar of visual design*, 2 edn, Routledge, London and New York.

- Krippendorff, K. (2004), *Content analysis: an introduction to its methodology*, 2 edn, Sage, London and Thousand Oaks, CA.
- Ledin, P. and Machin, D. (2018), *Doing Visual Analysis: From Theory to Practice*, Sage, London.
- Levshina, N. (2015), *How to do Linguistics with R. Data exploration and statistical analysis*, John Benjamins, Amsterdam/Philadelphia.
- Manovich, L. (2012), How to compare one million images, in D. M. Berry, ed., ‘Understanding digital humanities’, Palgrave Macmillan, Basingstoke, pp. 249–278.
- McElreath, R. (2016), *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, CRC Press, Boca Raton, FL.
- R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: <https://www.R-project.org/>
- Saldaña, J. (2016), *The Coding Manual for Qualitative Researchers*, 3 edn, Sage, London.
- Schoonenboom, J. and Johnson, R. B. (2017), ‘How to Construct a Mixed Methods Research Design’, *Kölner Zeitschrift für Soziologie und Sozialpsychologie* **69**(Suppl 2), 107–131.
- Teddlie, C. and Tashakkori, A. (2011), Mixed Methods Research: Contemporary Issues in an Emerging Field, in N. K. Denzin and Y. S. Lincoln, eds, ‘The Sage handbook of qualitative research’, 4 edn, Sage, Los Angeles, pp. 285–299.
- Thomas, M. (2014), ‘Evidence and circularity in multimodal discourse analysis’, *Visual Communication* **13**(2), 163–189.
URL: <http://vcj.sagepub.com/content/13/2/163.abstract>
- Winter, B. (2020), *Statistics for Linguists. An Introduction Using R*, Routledge, New York and London.