PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

David J. Hand
Department of Mathematics and Institute for Mathematical Sciences,
Imperial College, London

SIZE MATTERS: MEASUREMENT AND SCIENCE

1. Introduction

I want to begin with the observation that we are surrounded by measurement. Indeed, not only are measurements all about us, but our view of the world is defined in terms of measurements. I sometimes describe this situation by saying that we *see the world through the spectacles of quantification*. By this I mean that our interpretation of the things we see is based on notions of quantity.

The ubiquity of measurement is indicated by even a casual everyday conversation. We speak of the cubic capacity of a car engine, volume or weight of ingredients in cooking, the times that athletes take to run a race, the distance of a star from the solar system, the proportion of the population who voted for a particular candidate, the Gross Domestic Product of the economy, and so on. We even speak of the intensity of feelings – and that example might raise the question of whether notions of quantification are being stretched beyond their breaking point. I shall return to this issue later.

There are several reasons for the ubiquity of measurements. Measurements – numerical values assigned to an attribute to indicate its magnitude – enable unambiguous communication. If I determine that the height attribute of a bridge is 6 metres, and I tell you this, then you know exactly how high the bridge is. More than this, you know what will fit under the bridge. That is, you can base calculations and conclusions on your knowledge of the height of the bridge. In general, this means that you can do engineering and organise businesses. Through such means you can build machines that will run and determine whether your company is earning sufficient money to stay afloat. Measurements have a universality that other ways of indicating magnitudes do not have. A child who tells you that a man is *tall* may mean

something quite different from an Olympic basketball player who tells you the same thing. But if either tells you a man is seven feet tall you know exactly what is meant.

It was not always like this, and units, of length, weight, or whatever must have a common definition to be useful – even old religious books such as the Bible and the Koran make this point. Step back to the dawn of civilisation and it is no wonder that different peoples and different cultures used different measurements for the same magnitude of the same attribute – they had no way of communicating with each other. Indeed, step forward right to the eighteenth century in France and you still find around 800 different names for length [Alder 1995], with much the same being true for other countries. Step forward to as recently as 1837 and you find (after an abortive earlier attempt in 1795) France adopting a nationwide 'metric system'. In 1875 a step towards an international system was made, with the establishment of the *Conférence Générale des Poids et Mesures*, involving 17 nations, holding its first meeting in 1889. And this march continues. Even now, though, it is far from complete across even the Western world: the United States clings on to road distances in miles, and in 2002 the UK witnessed a celebrated court case in which market traders who wanted to continue weighing out goods in Imperial units of pounds and ounces lost their appeals against European laws requiring them to use metric measures.

This tendency towards global units (one might even refer to it as an aspect of 'globalisation') was inevitable with ever more advanced transport and communication technologies. Such changes meant that people from distant places could trade and talk to each other. To do that, they needed universal terms: they needed to know exactly what they were getting when they bought a barrel of grain or a length of cloth. They needed to know that conclusions reached by calculating using one system of measurement would not be different from conclusions reached using another system. If I measure the cloth in inches, determining how much I need for the suits I intend to make, I need to know that I will still have enough cloth even if *you* measure it in centimetres. I need a way of converting from inches to centimetres and if I can do that I effectively have a common unit, be it inches or centimetres.

In parallel with this march of quantification towards the use of common units, there has also been a march of quantification in terms of what can be measured. This, too, has a history stretching back to the dawn of time and this, too, is continuing. Amongst the earliest things to be measured were length and weight. Such measurements were needed for constructing the pyramids and stonehenge. An obvious approach is to compare the length to be measured with a standard 'unit' length. It is not so obvious how to measure area. Early definitions included the time it took to walk around an area. No doubt the reader can spot the weakness in this approach! Other measures were based on how long the area took to plough – an improvement over the walking round method, but hardly very accurate. Volume also had its own problems: was the volume of grain which a container could hold heaped or flat to

the surface? What if it settled? An interesting aspect of all of this is that it led to different names for measurements of what we would now regard as the same attribute associated with different kinds of things: different measures for volume of fish and grain, different measures for weight of diamonds (carats) and people (kilograms). Sometimes this lingers on: in English the height of horses (surely a 'length' measurement) is in terms of 'hands', not inches or centimetres; the depth of water (surely again a 'length' measurement) is measured in fathoms. And sometimes entirely different units are introduced for the same attribute for other reasons: we do not state the distance to stars in terms of metres, but rather in term of light years. In any case, no-one nowadays would think twice about the difficulties of measuring length, weight, area, and volume. Such measurements are part of our conceptual infrastructure for the world around us.

Temperature measurement is also part of our conceptual infrastructure. We not only speak of something being hotter than something else, but we can easily, indeed effortlessly determine the numerical value of the temperature of, for example, a patient in a hospital using a digital thermometer which gives a direct reading. But temperature is not so straightforward a concept as length or weight. While I can see how many foot rules it takes to reach the height of a seven foot tall man, it is not so obvious to see how many 'one degree' hot objects I can place to determine the man's temperature, nor where I would place them. Clearly something deeper is needed in order to measure temperature. Indeed, something rather deeper is needed in order to *understand* temperature. This extra depth is illustrated by the fact that early thermometers were called 'thermoscopes', devices for informal viewing of temperature, like microscopes and telescopes, rather than devices for assigning a numerical value to it. And this leads to a very important point. The determination of measurement procedures for attributes is intimately interwoven with defining those attributes - with working our what they really are. Temperature is different from heat, and a body can have a large heat content with a low temperature and vice versa (a beautiful history of the measurement of temperate is given by [Chang 2004]). Developing new measuring instruments is not simply an arena for craftsmen, but may be a core part of constructing the theoretical understanding of the concepts.

If temperature is a difficult concept to define, and if its definition and measurement are closely interwoven, then how much more this is true for concepts such as 'attitude' (say, towards a political stance), intelligence, pain, and economic growth. But all of these things are now measurable. Developing ways to construct measuring instruments for such concepts has been part of deciding what those concepts 'really are'. The definition and the measurement are two sides of the same coin. As I said at the start, we are surrounded by concepts of measurement, and only some of these are accessible by simple 'direct' measurement like length and weight. In the next section I discuss how we get from measuring things such as length to measuring things such as attitude, intelligence, pain, and economic growth.

2. What is measurement

I hinted above that certain kinds of measurement were relatively straightforward, and indeed have been with us for thousands of years. These include the measurement of length and weight. The basic structure of such measurement systems is to define a mapping from the relationships between a set of objects A in the physical world to corresponding relationships between numbers in a set R. Thus if $a,b \in A$, with A a collection of pebbles, we can map a and b to numbers $x(a)$ and $x(b)$, such that $x(a) > x(b)$ whenever an object a extends a spring balance further than object b (which we denote as $a \succ b$; that is, whenever a is heavier than b). Moreover, if $a,b,c \in A$, we can map the triple of objects a, b, and c to the triple of numbers $x(a)$, $x(b)$, and $x(c)$ such that if two of the objects (a and b, say) balance the third, c, on a weighing scales (which we denote as $a \circ b = c$) then $x(a) + x(b) = x(c)$.

In general, a collection of objects may have various attributes we wish to represent numerically. Different attributes will have different relationships which we wish to represent. For example, for the weight attribute of the pebbles we wished to represent the $\succ$ and $\circ$ relationships but we might measure the hardness attribute by seeing which pebbles would scratch which others (harder will scratch softer, but not vice versa), denoted $\succ$, and mapping this to the numerical order relationship $>$.

These ideas do not completely settle measurement issues, even for systems as simple as length and weight. For example, in place of addition in the above, we could have chosen to map to multiplication: the balance of the three objects a, b, and c in the ternary relation $a \circ b = c$ being mapped to numbers $y(a)$, $y(b)$, and $y(c)$ such that $y(a) \cdot y(b) = y(c)$. Obviously, in such cases there must be an isomorphism between the $x(\cdot)$ and the $y(\cdot)$ since they are representing the same physical system. Put another way, there must be a mapping between the two sets of numbers, since they both provide valid representations of the physical system. (In this case the mapping is given by log and its inverse.) Such mappings are called 'admissible' or 'permissible' mappings, and an even more straightforward example of such a mapping is a change of unit in length or weight measurement. This preserves, as addition, the numerical operation to which the physical system is being mapped. Non-admissible transformations are transformations outside this special class for a particular system. If the representing numbers are subjected to such a transformation they will no longer represent the physical system - and contradictions will arise. A simple illustration of this can be seen from the following quotation, which appeared in the London *Times*: 'Temperatures in London were still three times the February average at 55°F (13°C) yesterday.' This statement is perfectly fine, but it does prompt the question - so what is the February average? It se-

ems that the answer might be a third of 55°F, which is 18,3 °F, or it might be a third of 13°C, which is 4,3 °C. Unfortunately, the first of these is below freezing while the second is above freezing. Clearly something is wrong, and what is wrong is that rescaling is not an admissible transformation for switching between °F and °C as numerical representations of temperature.

I illustrated how the relationship between objects in terms of a particular attribute could be mapped to different equally valid numerical relationships using addition and multiplication. But different relationships between objects can also both be mapped to the same numerical relationship. For example, we can map end-to-end concatenation of rulers (that is, placing them end to end in a straight line) to addition, so that the end-to-end concatenation relationship $a \circ b = c$ maps to $x(a) + x(b) = x(c)$, in the perfectly familiar way of representing the length of objects. But we can also place rulers end-to-end at 90° (representing this way of combining two rulers a and b, by $a * b$) and assign numbers $x(a)$ and $x(b)$ to them so that the length $x(c)$ of the hypotenuse of the resulting triangle corresponds to an object c. That is, $a * b = c$ maps to $x(a) + x(b) = x(c)$.

Since different physical ways of combining objects can be mapped to the same numerical operation, and a single physical way of combining objects can be mapped to different numerical operations, there is a real possibility of confusion. And, indeed, such confusion does arise. Hand [2004] gives two examples: the controversy in the world of psychophysiology arising from the fact that the researchers drew different conclusions according to whether they measured skin resistance or skin conductance; and the apparently contradictory results of the impact of life events and a close partner on the development of depression.

Indeed, although all of the above examples mapping to addition have involved some sort of concatenation operation, this is not essential. *Conjoint* measurement is also a mapping to addition, but this requires only that objects can be rank ordered - there is no notion of concatenation. Further details are given in Gustafsson *et al* [2000].

Despite their complications, all of the above measurement procedures are relatively straightforward. We looked at an empirical system (which may be a physical system, but could be psychological, economic, social, etc) and found a numerical representation which preserved the relationships in that empirical system by relationships between numbers. Unfortunately, many of the systems we would like to measure defy simply numerical mappings. Intelligence provides a nice example. There is no simple concatenation, or even simple way of ordering people, which permits us assign numerical scores. We need a more elaborate measurement procedure. Inflation rate provides another example. The idea that different degrees of inflation might be empirically concatenated in some way seems meaningless.

Our aim, in measurement, is to assign numbers to represent the magnitude of the attribute in question. The lack of a simple empirical way of comparing and combining the objects in terms of the attribute suggests that either we will need to seek a more elaborate way in which to assign the numbers or we will need to restrict things so that a simple way emerges. Both approaches are widely used.

The first approach is achieved by latent variable models. Based on a theoretical understanding of the phenomenon being studied, one develops a model for the relationship between the (unobserved, latent) attribute one wants to measure and other attributes (the 'manifest variables') which are more amenable to direct measurement. Suitable statistical and mathematical modelling tools are then applied to infer the value of the latent variable from the observed values of the manifest variables. Factor analysis is perhaps the most widely used example of such an approach. The measurement of intelligence is an important example of this type of measurement. 'Intelligence' per se, is neither directly measurable nor well-defined. But other things related to it are clearly defined and are can easily be measured: scores on simple reasoning tasks, on understanding, on symbol manipulation, for example. If we can construct a theory relating the underlying notion of intelligence to such things which we can measure, then we can attempt to infer the latent variable 'intelligence'. But note, and this is important, exactly what we mean by intelligence here, how we have *defined* intelligence, is implicit in the theory we have constructed and the manifest variables we have chosen to measure.

Scaling methods, such as Thurstone, Likert, Guttman, and Coombs scaling can be regarded as informal and rather ad hoc latent variable models. In a Likert scale, for example, each of a set of (manifest) items is scored on a simple numerical scale (perhaps containing just five ordered values) and these are summed to yield an overall score.

The second approach is more direct and simply specifies conditions which an object must satisfy in order to achieve a certain measurement score. For example, the Apgar score of the clinical condition of a newborn infant takes values 0, 1, or 2 for each of colour of complexion, heart rate, respiration rate, reflex response to nose catheter, and muscle tone, and sums these to yield a total score. Quality of life provides another example. I might decide that amount of pain, close friends, an active social life, and an ability to carry out ordinary daily activities are the key aspects of quality of life, and I might specify the weights to be assigned to each of these which will sum to give an overall quality of life score. By specifying things tightly in this way, I have produced a unique definition. It is well-defined, so we can communicate using it and explore its relationship to other variables. Of course, it is entirely possible that you might prefer a slightly different definition of quality of life. Provided your definition was also well-defined, there is no reason why we should not use that as an alternative. The point is that, as well as describing how quality of life is to be measured, we have also defined what we mean by quality of life: you would mean something slightly different from what I mean.

Both approaches introduce extra information not apparent in the original empirical system. I term this extra information *pragmatic* information. It is introduced on practical grounds, or for convenience, or for other reasons, and serves to help to define precisely what it is one is seeking to measure while simultaneously specifying how that attribute is to be measured. In the first (latent variable) case the pragmatic information is the theoretical construct on which the attribute is based, as well as the choice of manifest variables. In the second case it is the definition of what is to be included in the construction of the measurement instrument and how it is to be included.

All measurement is thus a combination of representational and pragmatic aspects.

3. The power of measurement

When a measurement procedure has a strong representational component, the admissible transformations tell us what alternative numerical representations will provide other valid descriptions of the empirical system. Different classes of admissible transformations induce different 'scale types'. If the admissible class includes all monotonic transformations, then the scale is ordinal. If it includes affine transformations it is an interval scale. If it includes rescaling transformations it is a ratio scale. In fact, these scale types were proposed in the first half of the twentieth century by S.S. Stevens, but more recent work has formalised them in terms of the concepts of *homogeneity* and *uniqueness*, which characterise structure preserving transformations of the empirical system.

One of the consequences of a measurement representation having tightly constrained admissible transformations is that the form of scientific laws involving those measurements are also constrained. R.D. Luce proposed that for a law to be scientific it should have two important properties: (i) admissible transformations of independent variables should lead only to admissible transformations of the dependent variables; (ii) the mathematical structure of theories should be invariant to admissible transformations. Both of these properties seem eminently reasonable. Regarding (i), if I change the units of measurement of some ratio scale variable in a formula, I do not expect other variables to suddenly enter as squared terms or log transformations. Regarding (ii), if I change the units of measurement of a ratio scale variable in a formula, I do not expect addition to be replace by multiplication.

These proposed properties in fact turn out to be immensely powerful ideas for investigating proposed scientific relationships. For example, consider two physical attributes, $x(\cdot)$ and $y(\cdot)$, both measured on ratio scales (so that the admissible transformations are rescaling transformations – changes of units). Suppose we conjecture that x and y are related by some unknown function f: $y(a) = f(x(a))$.

Since we know that x and y have ratio scales, Luce's property (i) tells us that if we rescale x by a constant r (that is, if we change the units of measurement of x), then y must also be rescaled by some constant, $s = s(r)$, say. We thus have

$$s(r) f\left(x(a)\right) = f\left(rx(a)\right).$$

A little mathematics shows that the only functions f which satisfy this must have the form

$$f(x) = cx^b,$$

with b and c constants.

This is an extraordinary result. Purely from knowledge of the measurement properties of x and y we have deduced the form of their relationship. Of course, the numerical values of b and c have to be determined from experiment: from matching the formula to data from the real world.

These ideas have many extensions. They underlie the method of dimensional analysis, which is widely used in physics and other areas to explore the validity of proposed scientific relationships. A simple illustration is the following. A textbook suggested that the formula for the density function of the sample variance

$$s^2 = \sum (x - \bar{x})^2 / (n-1)$$

was

$$const \times \left(\frac{n-1}{2\sigma}\right)^{(n-1)/2} \left(s^2\right)^{(n-3)/2} \exp\left[-(n-1)s^2/\sigma^2\right].$$

However, a quick examination of the units in which each term in this expression are measured shows that its overall dimensionality is $D^{(n-5)/2}$, where D are the units in which x is measured. But this is supposed to be the pdf of a variance: its units should be D^{-2}. Something is wrong.

4. Measurement and statistics

The relationship between measurement and statistics has often been the source of much controversy. The argument is that certain kinds of statistical operations are not legitimate for certain kinds of measurements. A classic example of this is calculating the means of values which are from a purely ordinal scale. For example, suppose we have two sets, each of three objects, with values, on an ordinal scale, of $\{1,2,6\}$ and $\{3,4,5\}$. The mean of the first set is 3, which is less than that of the second set 4. However, arbitrary monotonic (increasing) transformations preserve the empirical content of ordinal scales - these are the admissible transformations. So an equally legitimate numerical representation of our six objects would map $\{1,2,3,4,5,6\}$ to $\{1,2,3,4,5,12\}$. Unfortunately, the means of the two sets of three

are now, respectively, 5 and 4. The structure preserving transformation has resulted in a different conclusion. Something is clearly wrong.

Early attempts to resolve this focused on restricting the kinds of statistical manipulations which could be applied according to the measurement scale. But things are not that straightforward. In a classic paper, Lord [1953] pointed out that 'the numbers don't remember where they came from' - the statistical operations make no explicit reference to how the numbers were derived. Section 2.5 of Hand [2004] describes the development of this controversy. Moreover, just because values are (for example) on a ratio scale does not mean that any statement involving means is sensible: a statement such as $\overline{x}.\overline{y} = \overline{z}$, with x, y, and z measured on ratio scales is not generally empirically meaningful.

A more elaborate resolution is needed. What is important is not the statistic relative to the measurement scale, but rather the use made of that statistic. It is the context in which a statistic is used which determines whether or not it is legitimate to use that statistic.

5. Conclusion

Notions of measurement pervade our languages and our lives, to such an extent that it would be impossible to function without them. Formal measurement procedures remove ambiguity, enable communication, and enhance understanding. Elementary physical measurement concepts were the earliest to be developed, with a direct and immediate link to the concept of counting: one counted the number of unit elements required to equal the thing to be measured (unit lengths, weights, etc.). But gradually, measurement procedures for more complex and abstract attributes have been developed. This development – this progress – has not been without its difficulties. The history of measurement shows that, at each step of the way, people have expressed reservations, even concern, about the attempt to objectify attributes they thought could not be measured. In physics, temperature is a good example. More recent ones include pain and happiness, attributes which require much more elaborate measurement procedures, and which have a much heavier pragmatic component than do things like length and weight. And, as measurement technology continues to progress, no doubt others will continue to express anxiety. The fact is, however, that, as Lord Kelvin put it in 1888: "When you can measure what you are speaking of and express it in terms of numbers, you know something about it. When you cannot express it in terms of numbers, your knowledge of it is of a meagre and unsatisfactory kind".

References

Alder K. (2002), *The Measure of All Things: the Seven-Year Odyssey that Transformed the World*, Little, Brown, London.

Chang H. (2004), *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press, Oxford.

Gustafsson A., Herrmann A., Huber F. (eds.) (2000), *Conjoint Measurement: Methods and Applications*, Springer-Verlag, Berlin.

Hand D.J. (2004), *Measurement Theory and Practice: the World Through Quantification*, Arnold, London.

Lord F.M. (1953), *On the Statistical Treatment of Football Numbers*, „American Psychologist" 8, s. 750-751.

WIELKOŚĆ MA ZNACZENIE: POMIAR I NAUKA

Streszczenie

Koncepcje pomiaru są tak wszechobecne, że często ich nie zauważamy: są to po prostu części otoczenia koncepcyjnego, w którym funkcjonujemy. Jednak nie zawsze tak było. Artykuł bada istotę pomiaru, pokazując jak skonstruowane są procedury pomiaru i jak się one ściśle przeplatają z teoriami naukowymi.