AMERICAN JOURNAL OF MATHEMATICAL AND MANAGEMENT SCIENCES

ORDERED CATEGORIES: GRAPHS SHOWING EMPIRICAL RELATIONS BETWEEN ONE PROPORTION AND ANOTHER, AND COMPARISON WITH THEORIES

T. P. Hutchinson
Centre for Automotive Safety Research
University of Adelaide
South Australia 5005
Australia

SYNOPTIC ABSTRACT

Some dependent variables are not fully quantitative and cannot be exactly measured, yet the categories are not merely qualitatively different states: rather, observations are classified into ordered grades. Are the proportions of observations in the several grades related to one another in some systematic way, for example, do they arise from an underlying normal distribution? In some datasets, there are just three grades, and comparison is made of several different circumstances. An important question is whether the proportion in category 3 is positively correlated with the ratio of the proportions in categories 2 and 1. Statistical models are given which permit methods of linearizing this relationship to be suggested. The areas of application considered are as follows: injury severity in road accidents; "don't know" responses in contingent valuation surveys; fracturing and shape of particles; taphonomy (i.e., the extent of deterioration of biological material over time); heart rate variability; pyrolysis (liquid may be intermediate between gas and solid); errors in neuropsychological tests (some may be less serious than others); errors in speechreading (one condition may be more difficult than another).

Key Words and Phrases: ordinal categories; linearizing transformations; P-P plots; injury severity; particle science; contingent valuation; "don't know" responses; taphonomy; heart rate variability; pyrolysis; cognitive neuropsychology; speechreading

2005, VOL. 25, NOS. 3 & 4, 365-399
0196-6324/05/030259-35 $40.00

1. INTRODUCTION

It is common in many fields of science to assign grades to some dependent variable, because it is impossible to measure it, and to analyse statistically those grades. Injury severity is one example. It is frequently graded as none, nonfatal, fatal. Typically, there is a hope that if we could measure the dependent variable, we would find there is only one variable to be analysed (that might be interpreted as the average or typical level), and that separate consideration is not needed for the different grades. Putting this another way, it should not matter (in the injury severity context) whether (a) we define a threshold between none and nonfatal, and consider the proportions above and below this (i.e., the proportion having any level of injury, fatal or not, is of interest), or (b) we define a threshold between nonfatal and fatal, and consider the proportions above and below this (i.e., the proportion of fatalities is what is of interest). Of course, if we really suspected that different variables were influencing fatal versus nonfatal injury from those influencing nonfatal versus no injury, we should be conducting separate analyses. (But with different processes operating, we might well completely avoid any notion of severity, and instead treat the various grades as qualitatively different.) Two questions then arise. Firstly, do the data support the idea that (a) and (b) give similar results? Secondly, can we find a theory to connect the two proportions? (Obviously, there will be a systematic difference between them; clearly, also, the relationship will depend upon how wide the middle category is; but what exactly might be the form of the relationship?)

The present paper will give some theory, and describe its application in a number of different areas of science. The background is that Hutchinson (1976a) considered data on injury severity in road accidents, the data being from two sets of

circumstances only and being detailed in the sense of having many (i.e., more than three) grades of severity, and then Hutchinson (1985) reviewed the use of similar statistical methods in other contexts. This line of work will be used to introduce the relevant concepts, in Section 2.1 below. With some variables, it is common to find them graded into only three categories (e.g., because the categorisation is used routinely, and anything more complex is considered impracticable). Hutchinson (1976b) discussed this for injury severity, and now the present paper reviews such datasets arising in a number of contexts.

The examples in this paper are as follows: injury severity in road accidents; "don't know" responses in contingent valuation surveys; fracturing and shape of particles; taphonomy (i.e., the extent of deterioration of biological material over time); heart rate variability; pyrolysis (liquid may be intermediate between gas and solid); errors in neuropsychological tests (some may be less serious than others); errors in speechreading (one condition may be more difficult than another). Examples 2 ("don't know" responses), 4 (taphonomy), and 7 (errors in neuropsychological tests) are published here for the first time, the others are summaries of earlier publications.

2. THEORY

2.1. Two sets of circumstances, many grades of severity

Let x be severity on an unobservable continuous scale of measurement, and $p_1(x)$ and $p_2(x)$ be the proportions of casualties having severity x or less in two sets of circumstances. The idea that it is unnecessary to give separate consideration to different grades of severity --- i.e., that there exists some definition of severity such that consideration of average severity is sufficient --- suggests that it is possible to represent the distribution of severity x in circumstances i as

$F(x-\mu_1)$, where the function F is the same in all circumstances. It is not necessarily the case that F is something simple and familiar, but it is worth investigating this. It might be, for example, that the distributions of severity in two circumstances are related to each other in the same way that two equal-variance normal distributions are. That is, the two distributions are $\Phi((x-\mu_1)/\sigma)$ and $\Phi((x-\mu_2)/\sigma)$ (where Φ is the conventional symbol for the normal cumulative distribution function, and σ is the standard deviation). Writing z for the inverse function of Φ (that is, $z(p)$ is the normal deviate corresponding to a cumulative proportion p), $z(p_1) = (x-\mu_1)/\sigma$ and $z(p_2) = (x-\mu_2)/\sigma$. Consequently, $z(p_2) = z(p_1) + \delta$, where $\delta = (\mu_1-\mu_2)/\sigma$, the difference between the means expressed in units of the common standard deviation.

What this tells us is to convert observed proportions p_1 and p_2 to the corresponding normal deviates and plot one against the other, with the data points referring to different severities, x; if the equal-variance normal model is valid, we will find a straight line whose slope is 1 and whose intercept is the difference between the means, expressed in units of the common standard deviation. (One could alternatively work with the quantities $1-p_1$ and $1-p_2$, the proportions having severity more than x, and convert these to the respective z-scores.)

Such graphs have been used in various fields of application (Hutchinson, 1985).

- In the general statistical literature, they are termed P-P plots (Gnanadesikan, 1977/1997, Section 6.2).
- In the context of signal detection theory (used originally in the psychology of perception), they are termed ROC (Receiver Operating Characteristic) curves. See, for example, Hutchinson (1981) and Swets, Dawes, and Monahan (2000) for reviews of the usefulness of these ideas outside of the original context of detecting a faint signal in a noisy background.
- In evaluating the accuracy of diagnostic tests, many medical papers these days provide a graph showing the relation between sensitivity (probability of a positive result when disease is present) and specificity (probability of a negative result in healthy people).
- In life testing of some item of equipment, the distributions of times before failure in different conditions are of interest. The expression of one distribution in terms of the other is sometimes termed an acceleration function.

2.2. Many sets of circumstances, three ordinal categories

Now suppose there are only three ordinal categories, but the proportions of these are observed in many (i.e., more than two) sets of circumstances. In the injury severity context, the ordered categories might be fatal, nonfatal, and none; or severe (including fatal), moderate, and slight (including none). Let q_2 be the proportion of casualties in the most severe category, and q_1 be the proportion in the middle category. (The subscripts on the q's refer to different categories; in Section 2.1 above, the subscripts on the p's referred to different sets of circumstances.) Then, if the equal-variance normal model is valid, q_1 and q_2 are related to two thresholds x_1 and x_2 by $q_2 = 1-\Phi((x_2-\mu)/\sigma)$ and $q_1+q_2 = 1 - \Phi((x_1-\mu)/\sigma)$. (For brevity, the subscript i has been omitted from μ.) Thus $z(1-q_2) = (x_2-\mu)/\sigma$ and $z(1-q_1-q_2) = (x_1-\mu)/\sigma$. Consequently, since $-z(1-p)$ may alternatively be written as $z(p)$, $z(q_1+q_2) = z(q_2) + \alpha$, where $\alpha = (x_2-x_1)/\sigma$, the difference between the two thresholds expressed in units of the standard deviation of the distributions.

What this tells us is to convert observed proportions q_2 and q_1+q_2 to the corresponding normal deviates and plot one against

the other, with the data points referring to the different sets of circumstances; if the model is valid, we will find a straight line whose slope is 1 and whose intercept is the difference between the thresholds, expressed in units of the standard deviation. A straight line with a slope other than 1 is also interpretable, see Section 2.4 below.

There may be some reason of interpretability or a convention in a particular topic area that lead to the use of some variation on this method of plotting.

- Perhaps the simplest variation is to work from the left hand end of the distribution instead of the right, i.e., with the quantities $q_0 = 1-q_1-q_2$ (the proportion in the least severe category) and $q_0+q_1 = 1-q_2$ (the proportion having the lowest or the middle severity). Another choice is to plot q_2 versus q_0. Since $q_0 = 1-(q_1+q_2)$, this is equivalent to plotting q_2 versus q_1+q_2, but with one of the axes being reversed.
- If the quantities plotted are q_2 and q_1+q_2 (or q_0 and q_0+q_1), the proportion in the most extreme category is both plotted horizontally and is included in what is plotted vertically; consequently, there may be concern that an observed correlation is in some sense artificial. This concern may be unjustified, as when q_2 is much smaller than q_1+q_2, which it typically is when we are referring to severities of injury. Even so, one might choose q_2 as the horizontal axis and plot q_1 vertically. The disadvantage of this is that the expected relationship is nonmonotonic: when q_2 is close to 0, a positive association with q_1 is expected, but when q_2 is close to 1, a negative association with q_1 (and a positive association between q_0 and q_1) is expected. Consequently, one might choose q_2 as the horizontal axis, and plot $q_1/(1-q_2)$ vertically. For example, it is expected that there will be a positive association between the proportion of fatalities and the proportion of nonfatally injured casualties who are seriously (rather than slightly) injured. There will be an example of this type in Section 9 below.

Figure 1(A) is a plot of q_1+q_2 versus q_2. The data are from Kidwell, Rothfus, and Best (2001), and will be discussed as example 4 below. In Figure 1(B), the z's have been plotted. A straight line of slope 1 appears to be an appropriate description. In Figure 1(C), the proportion q_1 is plotted directly against q_2.

The next two subsections discuss variations on this basic idea. Some readers may wish to go to the applications starting in Section 3, and refer back to 2.3 and 2.4 when needed.

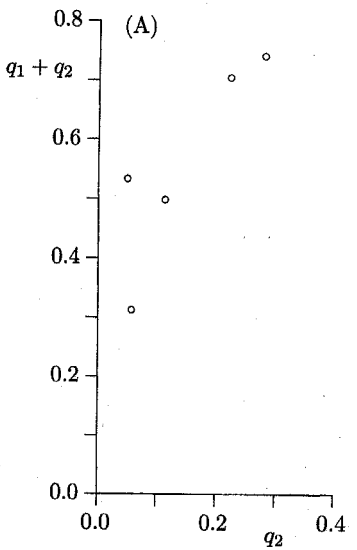

FIGURE 1. Data from Kidwell, Rothfus, and Best (2001, Fig. 11) on encrustation. (A) Relation between proportion of shells with high damage and proportion with either high or low damage.

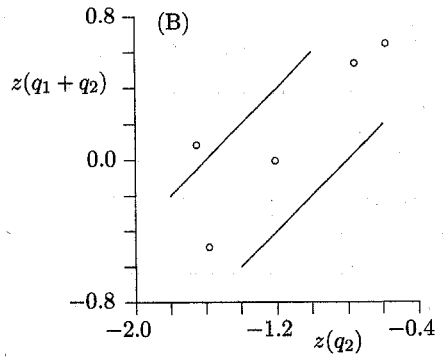

FIGURE 1 (continued). (B) The proportions have here been transformed to the corresponding normal deviates, $z(q_2)$ and $z(q_1+q_2)$; the straight lines are of slope 1.

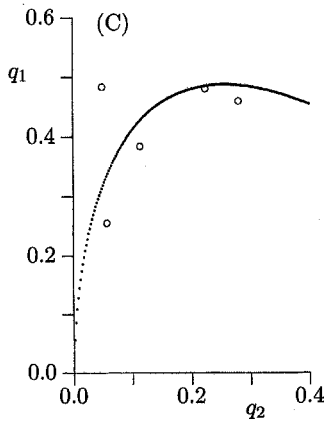

FIGURE 1 (continued). (C) Relation between proportion of shells with high damage and the proportion with low damage.

2.3. Varieties of one-parameter theory

The assumption that the F's are normal distributions led to a linear relation between the z-transformations of proportions. A linear relation between some more elementary transformations may be more appealing.

- *Logistic model*. Suppose the F's are logistic distributions with the same standard deviation. (Logistic distributions are similar, but not identical, in shape to normal distributions.) There is an elementary expression for the cumulative distribution function: $\exp(x-\mu)/[1+\exp(x-\mu)]$. Thus q_1 and q_2 are now related to two thresholds x_1 and x_2 by $q_2 = 1 - \{\exp(x_2-\mu)/[1+\exp(x_2-\mu)]\}$ and $q_1+q_2 = 1 - \{\exp(x_1-\mu)/[1+\exp(x_1-\mu)]\}$. The transformation we now want is $\ln[q/(1-q)]$, which is referred to as the logit of q. Then $\text{logit}(q_2) = \mu - x_2$ and $\text{logit}(q_1+q_2) = \mu - x_1$. Consequently, $\text{logit}(q_1+q_2) = \text{logit}(q_2) + (x_2 - x_1)$.
- *Exponential model*. In this example and the next, the F's do not have different location parameters μ, but have different scale parameters λ. Suppose the cumulative distribution of severity x in particular circumstances is $1 - \exp(-x/\lambda)$. Then $q_2 = \exp(-x_2/\lambda)$ and $q_1+q_2 = \exp(-x_1/\lambda)$. Consequently, q_1+q_2 is a power function of q_2, and $\ln(q_1+q_2) = (x_1/x_2).\ln(q_2)$.
- *Reversed exponential model*. The exponential distribution leads to a different relationship if the horizontal axis is reversed, with $x_2 < x_1$, and the most severe injury corresponding to values less than x_2, and the middle category of severity corresponding to values between x_2 and x_1. Now, $1 - q_2 = \exp(-x_2/\lambda)$ and $1 - (q_1+q_2) = \exp(-x_1/\lambda)$. Therefore, $1 - (q_1+q_2)$ is a power function of $1 - q_2$, and $\ln[1-(q_1+q_2)] = (x_1/x_2).\ln(1-q_2)$.

An attractive feature of each of these examples is that they can be linearised: in each case, some simple transformation W exists such that $W(q_1+q_2)$ is a linear function of $W(q_2)$. Hence graphical analysis of data is feasible.

Although an assumption about the F's enables us to derive a relation between q_2 and q_1+q_2, the reverse is not the case. That is, given a relation between q_2 and q_1+q_2, the class of F's is not determined. To illustrate this, consider the last of the above examples. (Remember, here q_2 is the proportion to the left of x_2, and q_1+q_2 is the proportion to the left of x_1.) For x being between $-\infty$ and ∞, let $F(x) = 1 - \exp[-\exp(x-\mu)]$ (this is one of the "extreme value" distributions). Then $1 - q_2 = \exp[-\exp(x_2-\mu)]$ and $1 - (q_1+q_2) = \exp[-\exp(x_1-\mu)]$. Consequently, $1 - (q_1+q_2)$ is a power function of $1 - q_2$, as before.

2.4. Two-parameter theories

The model based upon the normal distribution (Section 2.2) may be generalised to a two-parameter model (Hutchinson, 2002b, 2005). The motivation for doing this would be a finding of an approximately linear relationship between $z(q_1+q_2)$ and $z(q_2)$, but with a slope different from 1. Suppose the F's are normal, with standard deviation not constant but instead linearly related to the mean. Let the two thresholds be at x_1 and x_2, the mean of a distribution be μ and its standard deviation be $1 + b\mu$. Then $z(1-q_2) = -z(q_2) = (x_2 - \mu)/(1 + b\mu)$ and $z(1-q_1-q_2) = -z(q_1+q_2) = (x_1 - \mu)/(1 + b\mu)$. After a little algebra, we find $z(q_1+q_2) = A + B.z(q_2)$. In this equation, the intercept A is $(x_2-x_1)/(1+bx_2)$ (that is, the distance between the two thresholds, expressed in units of what the standard deviation would be if the mean of the distribution were at the higher threshold), and the slope B is $(1+bx_1)/(1+bx_2)$ (that is, the ratio of what the standard deviation would be if the mean of the distribution were at the lower threshold to what it would be if the mean of the distribution were at the higher threshold).

The same method of generalisation may be employed with the models in Section 2.3. Consider a location-scale family of distributions, $F((x-\mu)/\beta)$, with β being linearly related to μ. Starting from $q_2 = 1 - F((x_2-\mu)/(1+b\mu))$ and $q_1+q_2 = 1 - F((x_1-\mu)/(1+b\mu))$, we can readily a deduce a linear relationship between $F^{-1}(1-q_2)$ and $F^{-1}(1-(q_1+q_2))$. In attempting to linearise an empirical relationship between q_2 and q_1+q_2, therefore, we might try any transformation that corresponds to the inverse function of a location-scale family of cumulative distributions (preferably a reasonably well-known one). See also Section 9.3, where there will be a particular reason for considering the exponential model as the baseline: seeking an improvement on this, a linear relationship between ln(-ln) transformations of probabilities is predicted, with a slope of 1 corresponding to the exponential model itself.

3. EXAMPLE 1: INJURY SEVERITY

3.1. Empirical evidence

Empirical evidence that there is a positive correlation between the proportion of casualties killed and the proportion "seriously" injured was reported by Hutchinson (1976b). The examples included:

- For some hours of the day, a high proportion of pedestrian casualties are killed and a high proportion seriously injured, while at other hours these proportions are both relatively low.
- For some age groups, a high proportion of pedestrian casualties are killed and a high proportion seriously

injured, while for other age groups these proportions are both relatively low.

But this correlation did not appear in all circumstances. For example, when considering pedestrians struck by different types of vehicle, there was no correlation.

Of course, there is good reason to expect a positive correlation. Many of the variables that distinguish one set of circumstances from another would be expected to be related to either the violence of the impact (speed, or the acceleration inflicted), or to the victim's susceptibility. Violence and susceptibility are each thought of as wide-acting variables, likely to have an effect through the whole range of injury severity, affecting the proportions killed and seriously injured in similar ways.

Figure 2 refers to pedestrian casualties aged 50 or more, injured on roads where the speed limit was 60 km/h or less, in South Australia in the period 1980-2004. The points plotted are for different age groups, with casualties in their 50's at upper right and casualties in their 80's at lower left. The relationship seems to be a straight line of slope 1, as it was in Figure 1(B).

3.2. Allowing for differences in assessment of severity

Different people, or different organizations, are likely to differ slightly in their interpretations of rules governing the classification of severity of injury. An analysis that ignored these differences would be less sensitive than one that took them into account. Hutchinson and Lai (1981) examined the severity of single-vehicle non-pedestrian crashes in five police force areas of the U.K., and found that the police forces did differ.

3.3. Fatalities are important, but rare: Appropriate weighting

Fatalities, fortunately, are relatively rare. Consequently, there are many road safety studies in which random variation is a

large contributor to the number of fatalities. Suppose the numbers of casualties in the table below were observed.

	Slight	Serious	Fatal
Before	70	28	2
After	80	16	4

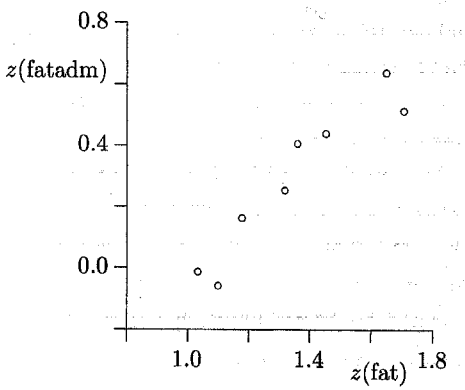

FIGURE 2. Proportion of fatalities and proportion of either fatal or hospital-admitted, in each case transformed to the normal deviate.

If some measure has been introduced aimed at reducing severity, these data suggest a degree of success. The proportion of serious and fatal together has fallen from 30 per cent to 20 per cent, and the increase in the fatalities might be ignored, on the grounds that the numbers are obviously too few to give weight to. But alternatively, the numbers might be taken at face value, in which case it is very questionable whether the After situation is to be preferred to the Before. The theory above offers a compromise between these alternatives. Experience with such data may have established that proportions of different injury severities typically behave as if derived from underlying normal distributions. This model could then be fitted to the data,

resulting in estimates of the means μ_1 and μ_2 in the two conditions. These would answer the question of whether the After situation was an improvement on the Before. The proportions of fatal and serious injuries under this model could also be obtained, and might be preferred over the raw numbers for input into an economic analysis.

Following this method would have the advantage of giving little weight to low numbers likely to be unreliable, without discarding them completely. The price that is paid is that it is necessary to make an assumption about how the data are appropriately modelled. (The result would also be subject to the limitation that a particular criterion for estimating the parameters, e.g., maximum likelihood, needed to be chosen.) It might be thought that the separation α between the thresholds for serious and fatal injury would be known through experience (and thus the task of estimating μ_1 and μ_2 would be easier). Unfortunately, this would probably not be the case: α is expressed in units of the standard deviation of the severity distributions, and this would change from one table of data to another, depending upon the degree of disaggregation of the data. (But having four or more severities of injury would be an improvement. Four severity categories would mean three thresholds separating them, and thus two separations α_1 and α_2. The ratio of these should be known through experience, as the effect of standard deviation would cancel out.)

3.4. Remark on failure of model

It might be noted that it is possible that the stiffness of a given thickness of an energy-absorbing material covering a hard surface is a practically-important variable for which the model of Sections 2.2-2.3 may be unsuitable: stiffness may have different effects at the high and low ends of the severity range. Low stiffness may be better for most impacts, but a poor choice

for high speeds, as the human who strikes it uses up all the available distance and then strikes the hard surface underneath. Thus there may be circumstances where it would be desirable for a test that two or more distributions of injury severity are the same to be sensitive to changes of both location and variability. References to nonparametric tests of this type are given by Hutchinson (2002a).

4. EXAMPLE 2: "DON'T KNOW" IN CONTINGENT VALUATION SURVEYS

The responses yes, no, and "don't know" may be available to survey respondents. A particular type of survey seeks to establish how much people would be prepared to pay (e.g., through taxation) for something good such as an environmental clean-up, or how much they would need to receive to compensate for something bad such as a noisy highway; the term contingent valuation is often used for this. Attempts are made to establish a value of life or a cost of injury by this method. Figure 3 illustrates the data in Table I of Wang (1997), the points referring to different amounts of money in an otherwise similarly-worded question. This question concerned how the respondent would vote if there were a referendum on an environmental improvement accompanied by a levy of a stated amount. A relatively low proportion were in favour when the proposed levy was high, and a relatively high proportion when it was low. The straightforward interpretation of "don't know" responses is as being intermediate between yes and no, like nonfatal injury is intermediate between fatal and no injury. Figure 3 shows how the proportion in favour and the proportion either in favour or unsure co-vary. In contrast to Figures 1(B) and 2, the slope seems a little less than 1.

The idea here is that "don't know" is an opinion in between yes and no, and the three proportions were co-varying in a sensible fashion. However, the proportions unsure were quite similar for all four proposed levies. In fairness, then, it

should be added that the data are also consistent with the hypothesis that the population includes a group of people who are uninterested or hostile or uncomprehending.

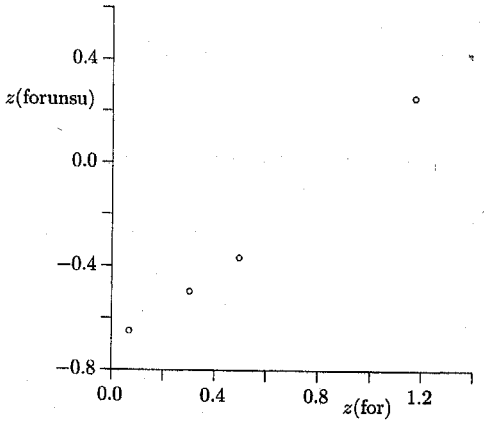

FIGURE 3. Proportion for and proportion either for or unsure about an environmental improvement plan, in each case transformed to the normal deviate.

5. EXAMPLE 3: FRACTURING AND SHAPE OF PARTICLES

This example, also, has some connection with traffic safety, in that the data concern some physical properties of gravels used in road construction. But it really comes within the field of particle science. Particulate material has various properties, such as the proportion of particles that are fractured, or the proportion that are flat or elongated, or the proportion that are both flat and elongated. There is room for debate on how fractured a particle must be before it is put into the fractured category, or how flat it must be before it is put into the flat category. Suppose we want to know how fractured are the particles of a gravel. We might determine the proportion of particles

having two or more fractured faces. A less stringent criterion is the proportion of particles having one or more fractured faces. We hope that both criteria reflect the same concept. Figure 3 of Hossain, Parker, and Kandhal (2000) (HPK) shows the relation between these two criteria for 24 chert and quartz gravels, and demonstrates that they are indeed highly correlated. Similar issues arise with other variables: in the same paper, the proportions of flat or elongated particles assessed according to 5:1 and 3:1 criteria were compared, and the proportions of flat and elongated particles assessed according to 5:1 and 3:1 criteria were compared.

What relationship might exist between proportions satisfying criteria of different stringency? A really good theory would connect the statistical variation in, say, degree of fracturing to the properties of the material (strength, brittleness, etc.) and the characteristics of the crushing process. But that seems beyond present-day knowledge. The lines fitted in HPK go to the other extreme of being purely empirical (a straight line in their Figure 3 and quadratic curves in their Figures 6 and 7).

The approach in Section 2 above seems applicable. HPK comment on the subjectivity involved in classification: concerning fracturing, they say that fractured faces need to be distinguished from chipped or flaked surfaces, and concerning shape, they say the positioning of a particle in the proportional callipers needs the exercise of judgment. They make the distinction between a definition in a test procedure being precise, and the use of that definition being imprecise. The equations in Section 2 are a step more theoretically justified than those used by HPK, and the parameters fitted therefore more interpretable. Hutchinson (2002b) demonstrated their relevance to the data in the paper of HPK; the specific approach was that of a family of normal distributions with standard deviation linearly related to the mean (see Section 2.4), as the z-transformations

of the proportions seemed to lie on a straight line whose slope was not 1.

6. EXAMPLE 4: TAPHONOMY

Taphonomy, according to the Oxford English Dictionary, is the branch of palaeontology that deals with the processes of fossilisation. (Also, research dealing with rotting over periods of months and years, which may be of forensic interest, or over periods of hundreds of years, which may be of archaeological interest, is carried out, as well as that concerning much longer periods.)

Taphonomists may wish to use death assemblages to make inferences about the environment thousands or millions of years ago, and thus an understanding of how biological material changes after death will be important. Consequently, they need to be able to assign grades of postmortem damage, and to analyse statistically those grades. Kidwell, Rothfus, and Best (2001) (KRB) presented data on the degree of damage to shells, graded as none, low, or high. They reported results for several different forms of damage. One of these was encrustation, see Figure 1(A). A high correlation can be seen: sites that have a high proportion q_2 of shells in the high damage category tend also to have a high proportion q_1+q_2 in the high plus low categories combined (and therefore a low proportion in the no damage category). As mentioned earlier, Figure 1(B) shows the proportions transformed to their respective z's, and that the relationship is now approximately a straight line of slope 1. Figure 1(C) shows q_1 plotted against q_2; the curve is the prediction made by the equal-variance normal model with a difference of 1.31 between the thresholds. Agreement between the curve and the data points appears satisfactory.

There is no guarantee that sites having a high proportion of shells in the high damage category will tend also to have a high proportion in the high plus low categories combined. It is an empirical question. And if this is the qualitative finding, it is also an empirical question whether changing the mean (but not the standard deviation) of a normal distribution is an appropriate quantitative description. The dataset in Figure 1 was selected for presentation because the method is successful. Figure 4 is similar to Figure 1(A), except that the nature of damage is fine-scale surface alteration. There appears to be no correlation. So the following questions are open. Is it usually the case that sites having a high proportion of shells in the high damage category tend also to have a high proportion in the high plus low categories combined? (If not, then damage should probably not be regarded as semi-quantitative, and instead one should concentrate on the qualitative features.) If it is sometimes the case, then in what circumstances (or, for what forms of damage) is it true and in what circumstances is it not? And, when it is the case, what quantitative form should the theory take --- the equal-variance normal model, or some other?

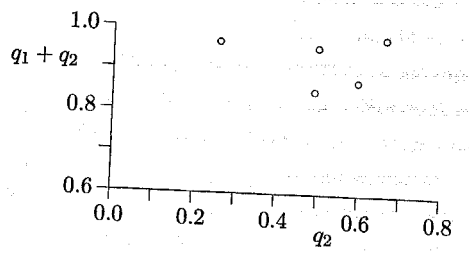

FIGURE 4. Data from Kidwell, Rothfus, and Best (2001, Fig. 11) on fine-scale surface alteration: relation between proportion of shells with high damage and the proportion with either high or low damage.

Incidentally, taphonomic datasets are sometimes encountered in which there are many grades of damage, and thus can be plotted in the same manner as in Hutchinson (1976a, 1985). There is a dataset in Fürsich and Flessa (1987, Figure 15) that is suitable for this. If $z(p_2)$ is plotted versus $z(p_1)$, the two distributions being the preservation quality of *Chione* in the mid channel and the outer channel, the result is found to be approximately a straight line of slope 1. But again, the method is not always successful: if the two sites compared are the mid channel and the mid flat, the points do not appear to lie on a straight line of slope 1.

7. EXAMPLE 5: HEART RATE VARIABILITY

One of the approaches to the measurement of heart rate variability is to consider the absolute differences between successive normal sinus intervals and determine the percentage that are greater than some threshold value. Having introduced the subject in this way, the obvious question is what threshold? And then we realise that an important issue is whether the distributions of intervals constitute a single-parameter family, because if they do, we will easily be able to convert results obtained using one threshold to their equivalents if another threshold had been used, by using a formula similar to those in Section 2 above (and if they do not, we are faced with a much more complex problem altogether). The most commonly used threshold is 50 ms, and the statistic obtained is termed pNN50. It is interpreted as a measure of cardiac parasympathetic modulation. Some studies have used thresholds other than 50 ms. Mietus, Peng, Henry, Goldsmith, and Goldberger (2002) proposed that pNN20, the percentage exceeding 20 ms, is a better statistic. Their reason was that several groups, including

normal/congestive heart failure, sleeping/waking, and young/elderly, were distinguished more decisively (lower p-value) by pNN20 than by pNN50.

Hutchinson (2003) pointed out that the relation between the two statistics is nonlinear, and that it is possible to transform them in such a way that the relation is linear. Plotting pNN20 versus pNN50 using transformed scales is likely to be generally useful --- a scatterplot is a natural tool to investigate the relationship, and different groups of people may be identified using different symbols. In the absence of a theory about the connection between them, and with the relationship being non-linear, there is a danger that a scatterplot will be uninterpretable. But the methods of Section 2 above might together make clear whether there are any deviations from one single relationship. If it did emerge that different information is conveyed by pNN20 and pNN50, this could be followed up by more detailed study of the dependent variable (e.g., pNN10 and pNN100 might be considered also), and of the independent variables (e.g., the identification of the conditions in which pNN20 and pNN50 behave differently). A choice between the alternative statistics would then be necessary. For example, if it were found that variation in parasympathetic modulation of heart rate led to a consistent relation between pNN20 and pNN50, but some other factor distorted pNN50 while leaving pNN20 less affected, then there would be a reason for preferring pNN20 over pNN50 (assuming the aim was to reflect parasympathetic activity).

8. EXAMPLE 6: PYROLYSIS

Gas, liquid, solid. In a sense, these have a natural order, liquid is between gas and solid, but this is a less convincing ordering than when referring to a dimension like severity of

injury or variability in heart rate. Whether or not it is a useful conceptualisation can only be determined by experience.

Güllü (2003) reported on the products obtained from pyrolysis of four materials (hazelnut shell, tea factory waste, tobacco stalk, and yellow pine wood). In the study of pyrolysis, it is common for a broad classification of products as gas, liquid, or char to be of interest. (Detailed chemical analysis is also of interest, of course.) Table 2 of Güllü's paper gives the proportions of gas, liquid, and char obtained from four source materials at each of eight temperatures. Hutchinson (2005) considered whether the proportion of char can be predicted from the proportion of gas. Do circumstances that lead to a high proportion of gas tend also to lead to a high liquid to char ratio, and thus a low char proportion? This would be so if there is a single unobserved (latent) variable that independent variables like pyrolysis temperature and source of material determine, which is then manifested in the proportions of gas, liquid, and char. It would imply a single equation is sufficient for predicting the latent variable from the independent variables; going from the latent variable to the three proportions would involve no additional variables.

Hutchinson (2005) plotted the proportion of char versus the proportion of gas, and found that if all 32 data points are included, the relationship is only an approximate one: the correlation coefficient is -0.76. However, many different chemical reactions are taking place in pyrolysis. It might be considered appropriate to restrict consideration to the lowest temperatures in Güllü's experiments: at temperatures 675, 725, and 775 K, the reactions are exothermic, but these temperatures are not so high that vigorous secondary reactions occur. It then turns out that there is one single relation (approximately a straight line) that applies to three of the starting materials and all three temperatures. (Restricting attention to these nine data points, the correlation is -0.97.) Tea factory waste is

different from the other materials, giving more char (and less liquid) at a given proportion of gas. A family of normal distributions with standard deviation linearly related to the mean (see Section 2.4) was used in Hutchinson (2005), as a plot of the z-transformations of the proportions was a straight line with a slope not equal to 1.

A sceptic might object that what has been done is to restrict attention to three out of eight temperatures, and three out of four source materials, and it might be thought that if one searches among subsets of any 32 pairs of miscellaneous numbers, it is almost inevitable that a relationship will be found. However, two datasets in Demirbas (2002) show a similar relationship, which is some evidence that it is not entirely fortuitous.

9. EXAMPLE 7: ERRORS IN NEUROPSYCHOLOGICAL TESTS

One strategy of research in the field of cognitive neuropsychology centers on testing of patients with injuries or diseases of the brain, and then attempting to infer what has gone wrong with them, in the hope that this will throw light on the structure of mental processes in healthy people. The present example concerns one aspect of the paper by Shallice, Rumiati, and Zadini (2000) (SRZ). Figure 5, taken from Figure 3 of that paper, shows a positive relationship between the proportion of responses that were correct and the proportion of wrong responses that were of a particular type, namely, single phoneme/letter errors. The eleven data points refer to four patients in three conditions, with one combination missing. The three conditions were repetition, reading, and writing (of nonwords). It is plain that SRZ attach some importance to this relationship, and would like to have a theory for it: they discuss the finding at p. 534, there is further discussion at p. 540, and on their Figure 3

itself they include a curve (it is quadratic and it passes through (0, 0), but no justification is given for either of these features). An explanation for the empirical relationship in Figure 5 will be proposed below that in general terms is similar to earlier examples of this present paper. But an additional feature will be an explicit model for the cut-off.

9.1. Assumptions and argument

On two grounds, it seems reasonable to suppose that single phoneme/letter errors are more similar to correct responses than other errors are. (a) The other types of errors are double or complex errors. These terms appear to indicate an error that is larger or more serious. (b) The empirical relationship (Figure 5) between single phoneme/letter errors and correct responses is positive.

There is variability: the patient does not always respond correctly, or always make one type of error. The existence of variability suggests we need some random variable, with some distribution. SRZ (p. 540) refer to a "resource" being damaged, and to the four patients being on a "single dimension" of impairment. It might be, then, that how much of the resource is given to the task determines whether it is performed correctly or not: if it is sufficiently great, the response is correct; if it is less, there is a minor error; less still, and there is a more serious error. For example, the random variable might be time devoted to the task. If sufficiently long, the response is correct; less, and a single phoneme/letter error results; less still, and there is a double or complex error. A specific assumption about the probability of ending processing and giving a response (at any particular moment) may be suggested: this probability is constant (i.e., does not change as time passes). This would mean that the giving of a response is a Poisson process. The consequence will be that time (elapsing until

response) has an exponential distribution. That is, the emission of a response is unconnected with the progress of the processing.

9.2. Predicted relationship

The idea that the distribution of the time is exponential implies that $\exp(-t/\mu)$ is the probability that the time exceeds t, where μ is the mean. In other words, the rate of occurrence of the event that terminates processing and triggers response is $1/\mu$. Both the degree of impairment and the difficulty of the task are envisaged as having their effect via μ: the greater the impairment or the more difficult the task, the smaller is μ. A correct response is given if the time exceeds some threshold; without loss of generality, this can be taken to be at $t = 1$; a single phoneme/letter error is made if time is between some value T and 1; and a more serious error is made if time is less than T. Now, if we use the symbol q_2 for the probability of a correct response and q_1 for the probability of giving an error of the single phoneme/letter type, what is plotted in Figure 5 is the conditional probability $q_1/(1-q_2)$ versus q_2. The probability of a correct response is $q_2 = \exp(-1/\mu)$; also, $q_1+q_2 = \exp(-T/\mu)$; thus the vertical axis in Figure 5 is $q_1/(1-q_2) = [\exp(-T/\mu) - \exp(-1/\mu)]/[1 - \exp(-1/\mu)]$. Finally, the predicted relationship in Figure 5 is $q_1/(1-q_2) = (q_2^T - q_2)/(1 - q_2)$.

Thus it has been possible to find a simple, theoretically based, functional form for the association that is evident in Figure 5. It has only one parameter, T. For the lower line in Figure 5, T has been taken to be 0.26 (the mean of $\ln(q_1+q_2)/\ln q_2$). To my eyes, the curve appears too flat, but at least this approach has predicted a positive relationship. The fit is not so poor as to compel rejection of this theory. And, there is no other theory competing with it.

One way forward would be to dismiss the idea that data for all patients in all conditions should lie on the same curve.

Instead, we might say that we now have an interpretation for the quantity $\ln(q_1+q_2)/\ln q_2$ (that is, it is T), and calculate this for each data point, and look for patterns in the values for the four patients in three conditions. On doing this, the following was found.

- It seems that T is unaffected by whether the task was repetition, reading, or writing: no pattern was found.
- In each condition, patient LT has the highest value of T.

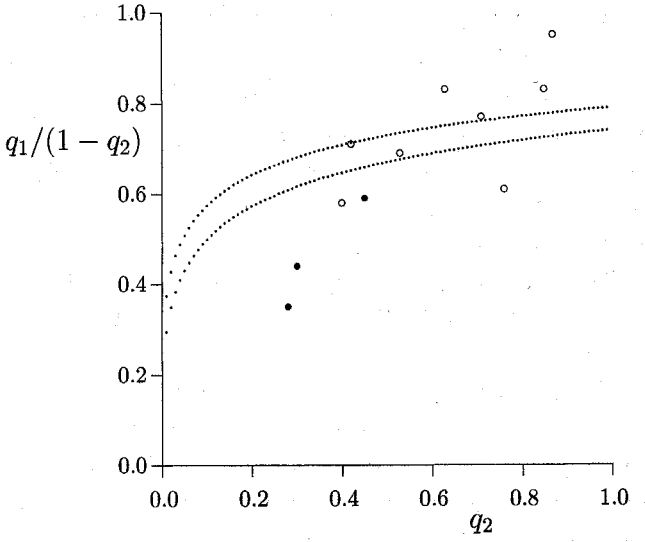

FIGURE 5. Proportion of single phoneme/letter errors $q_1/(1-q_2)$, as related to proportion of correct responses q_2: data for four patients in three conditions (Shallice, Rumiati, and Zadini, 2000). Solid disks refer to patient LT, open circles to the other three patients. Single-parameter curves having T = 0.26 (lower) and T = 0.21 (upper) are also shown.

SRZ regarded patient LT as being qualitatively similar to the other patients, but differing in degree, having suffered more

serious impairment. However, the fact that the values of T found in this patient are different from those of other patients could be taken as supporting an opposing view. (To say that LT is more severely impaired than the other patients is to say that his data points are on the left in Figure 5; to say that LT has a high T is a different issue --- given his values of q_2, his data points are lower on the Figure than would be expected.) Figure 5 has been drawn in such a way as to permit LT to be seen as different from the other patients. The upper line has been fitted with data from LT (shown as solid disks) excluded. This line has T = 0.21 and appears to be a good fit to the other eight points, with LT clearly deviating from the general trend. However, we would conclude that LT is qualitatively different from the other patients only if we found the one-parameter theory very plausible and persuasive. If, on the other hand, we feel it is permissible to adjust both the location and the slope of the relationship, we will see one single relationship in Figure 5. Two strategies for obtaining such a relationship, by modifying the theory so that two parameters can be chosen, thus obtaining a better fit for all of the data, will now be sketched.

9.3. Two-parameter theories

The first strategy adds a second level of randomness to the exponential distribution already present, and might take either of two forms.

- Let either a correct response be given (with probability c), or else the original theory apply (with probability 1-c). Then $q_2 = c + (1-c).\exp(-1/\mu)$, $q_1+q_2 = c + (1-c).\exp(-T/\mu)$, and an expression for $q_1/(1-q_2)$ in terms of q_2 could be obtained.
- Let the original theory apply, except that either response is instantaneous (the time elapsing is 0), and the probability of this is i, or else has an exponential

distribution (with probability 1-i). Then $q_2 = (1-i).\exp(-1/\mu)$, $q_1+q_2 = (1-i).\exp(-T/\mu)$, and an expression for $q_1/(1-q_2)$ in terms of q_2 could be obtained.

But why should c (or i) be the same for everyone, regardless of their degree of impairment, and in all conditions, regardless of difficulty? Until this can be answered, it is difficult to view c (or i) as anything more than a fudge factor to improve the fit of the prediction curve.

The second strategy involves introducing extra flexibility into the family of distributions, as in Section 2.4. Let us suppose the distribution of the resource is not exponential, but is normal. (Time devoted to the task could not have a normal distribution, because a negative time is impossible in this context, but the logarithm of time might have this distribution.) Furthermore, although the normal distributions still only vary in one parameter μ, extra flexibility is introduced by the standard deviation σ not being constant but being linearly related to the mean: $\sigma = 1 + b\mu$. The slope of the relationship is the extra flexibility. Without loss of generality, the threshold for a correct response can be taken to be 0. The lower threshold is T (negative). Then (as in Section 2.4) we find $z(q_1+q_2) = -T + (1+bT).z(q_2)$. Figure 6 shows the data from Figure 5 plotted in this way. T is estimated to be -0.92 and b is to be -0.15.

A similar idea leads to a direct generalisation of the original exponential model. Consider the cumulative distribution $1 - \exp[-\exp((x-\alpha)/\beta)]$. Again taking the threshold for a correct response to be 0, $q_2 = \exp[-\exp(-\alpha/\beta)]$, and $q_1+q_2 = \exp[-\exp((T-\alpha)/\beta)]$. If β is a constant, this may readily be seen to be indistinguishable from the exponential model: q_1+q_2 is a power function of q_2, as before (we have in effect transformed the time axis to the logarithm of time). But instead of β being constant, let us now say that β is linearly related to α. The consequence is that a linear relationship between transformations of q_1+q_2 and q_2

is predicted: $\ln[-\ln(q_1+q_2)] = T + (1+\gamma T).\ln(-\ln q_2)$. When the ln(-ln) transformation is used and the data plotted, a straight line is evident, similar to Figure 6.

This second strategy, whether it utilizes the normal distribution, the exponential distribution, or something else, is more attractive than the first, partly because of the convenient data processing (a linearizable relationship), and partly because the result is interpretable: it has permitted the conclusion that circumstances (severe impairment, or high difficulty of task) that are associated with a low probability of correctness are also associated with a high degree of variability in the latent random variable that reflects level of performance. (That is, low q_2 corresponds to low μ, and low μ is associated with high σ.)

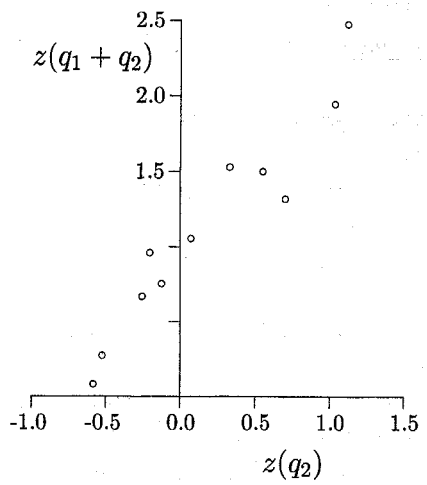

FIGURE 6. Data from Figure 5 with proportions q_2 and q_1+q_2 transformed to their corresponding normal deviates.

9.4. Concluding comment

This type of theory is sufficiently abstract that it may find application whenever there is a grading of the incorrectness of incorrect responses, as well as a distinction between correct and incorrect. The specifics of the theory are open to debate: it is mere speculation to say that the relevant resource is time; even if this were true, the question would arise whether the important thing is the passage of time *per se* or whether something has been utilized and exhausted during the passage of time; and any conclusion about the family of distributions (exponential, normal with constant standard deviation, normal with varying standard deviation, etc.) can only be tentative, on the limited evidence of four patients in three conditions.

10. EXAMPLE 8: ERRORS IN SPEECHREADING

This example concerns the proportion of errors made in speechreading, in two conditions. The transformation to a linear relationship is very similar to the previous examples. However, to consider the proportions being compared as differing in the location of a threshold may be considered a rather unnatural interpretation in this case.

When considering the probability of responding correctly to a test item, the following model is sometimes employed. There are quantities that might be called difficulty and ability, that respectively are characteristic of the item and the person responding to it. A person's ability varies from moment to moment, around some mean. A correct response is given if ability exceeds difficulty at the moment the item is attempted. Specifically, the logistic distribution is often used in this context. The expression for this (see Section 2.3) is $\exp(x-\mu)/[1 + \exp(x-\mu)]$. This expression represents the probability of ability being less than x, which is the probability of an incorrect response when item difficulty is x.

Now consider a comparison of two conditions. Let the difficulty in the first condition be d. The probability of correct response is $q_2 = 1 - \{\exp(d-\mu)/[1 + \exp(d-\mu)]\}$. In going to the second condition, suppose difficulty changes by an amount Δ, i.e., it becomes $d - \Delta$. Then the probability of correct response becomes $q_2' = 1 - \{\exp(d-\Delta-\mu)/[1 + \exp(d-\Delta-\mu)]\}$. Then $\text{logit}(q_2) = \mu - d$ and $\text{logit}(q_2') = \mu - (d - \Delta)$. Consequently, $\text{logit}(q_2') = \text{logit}(q_2) + \Delta$.

The practicalities of experimentation are likely to depart in two respects from what is implied in the previous paragraph. First, d was there viewed as a constant, but data will usually be based on a number of items of different difficulties. Second, each subject is likely to be presented with different items in the two conditions, in order to avoid the effect of memory. The experiment of Erber (1992), to be described in the next paragraph, had these features. Thus the random element, which was described above as moment-to-moment variation in a person's ability, also includes contributions from the differing difficulties of the items and from differing average difficulties of the two samples of items.

An experiment on speechreading of spoken sentences was reported by Erber (1992). The two conditions were presentation of the sentence in isolation, and presentation in response to a question. The 24 data points referred to different people. Erber found that subjects' probabilities of correctness tended to be higher for sentences that followed a question than for sentences presented alone. Hutchinson and Cairns (2000) showed, in their Figure 2, that the logits of the probabilities are approximately linearly related, with a slope of about 1. (They noted that their results were very similar whether the normal or the logistic distribution was employed.) Actually, much of the paper by

Hutchinson and Cairns was not focussed on the consistency or accuracy of the relationship between the logits, but was concerned with trying to explain the scatter in the relationship. That is, Δ was calculated for each subject, and then an attempt was made to relate Δ to characteristics of the subject: it was found that the extent of the advantage of the question-answer sequence tended to be less for older subjects.

11. DISCUSSION

If we see a scatterplot like Figure 1(A) or Figure 5, why might we care about fitting a theory to the data points?

- One simple answer is that if the relationship is considered interesting enough to show the scatterplot, surely a theory to explain the relationship is interesting too!
- We may wish to be able to convert results using one threshold to results using another, as with pNN20 and pNN50 (example 5).
- We might aim for a theory explaining the latent variable (e.g., in example 6, from the experimental conditions). It would then be easy to calculate the proportions in the several categories.
- We may wish to explore the limits of the relationship, and the implications of any failure to find the relationship.

In some contexts the proposed approach is very plausible. It would be surprising if it failed with severities of various diseases or types of damage. (The problem here is whether the concept of severity is unidimensional.) In other cases, it is much more questionable. For example, is liquid really intermediate between gas and solid? Is one type of response

really less correct than another, and less wrong than a third? In such contexts, the two probabilities are plotted in hope, rather than expectation, that a relationship will be found. If a relationship is found, this may be the start of a fruitful new line of enquiry.

ACKNOWLEDGMENTS

The Centre for Automotive Safety Research receives core funding from both the Department for Transport, Energy and Infrastructure (South Australia) and the Motor Accident Commission (South Australia). The views expressed in this report are those of the author and do not necessarily represent those of the University of Adelaide or the sponsoring organizations.

REFERENCES

Demirbas, A. (2002). Utilization of urban and pulping wastes to produce synthetic fuel via pyrolysis. *Energy Sources, 24*, 205-213.

Erber, N.P. (1992). Effects of a question-answer format on visual perception of sentences. *Journal of the Academy of Rehabilitative Audiology, 25*, 113-122.

Fürsich, F.T., & Flessa, K.W. (1987). Taphonomy of tidal flat molluscs in the northern Gulf of California: Paleoenvironmental analysis despite the perils of preservation. *Palaios, 2*, 543-559.

Gnanadesikan, R. (1977/2nd edition 1997). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Güllü, D. (2003). Effect of catalyst on yield of liquid products from biomass via pyrolysis. *Energy Sources, 25*, 753-765.

Hossain, M.S., Parker, F., & Kandhal, P.S. (2000). Comparison and evaluation of tests for coarse aggregate particle

shape, angularity, and surface texture. *Journal of Testing and Evaluation, 28*, 77-87.

Hutchinson, T.P. (1976a). Statistical aspects of injury severity. Part I: Comparison of two populations when there are several grades of injury. *Transportation Science, 10*, 269-284.

Hutchinson, T.P. (1976b). Statistical aspects of injury severity. Part II: The case of several populations but only three grades of injury. *Transportation Science, 10*, 285-299.

Hutchinson, T.P. (1981). A review of some unusual applications of signal detection theory. *Quality and Quantity, 15*, 71-98.

Hutchinson, T.P. (1985). Presenting one probability distribution as a function of another --- Some applications. *American Journal of Mathematical and Management Sciences, 5*, 103-123.

Hutchinson, T.P. (2002a). Should we routinely test for simultaneous location and scale changes? *Ergonomics, 45*, 248-251.

Hutchinson, T.P. (2002b). The relation between an extreme proportion and a less extreme proportion, in the context of the comparability of tests. *Journal of Testing and Evaluation*, 30, 255-257.

Hutchinson, T.P. (2003). Statistics and graphs for heart rate variability: pNN50 or pNN20? *Physiological Measurement, 24*, N9-N14.

Hutchinson, T.P. (2005). Proportions of gas, liquid, and char from pyrolysis of biomass. *Energy Sources, 27*, 1029-1034.

Hutchinson, T.P., & Cairns, D. (2000). Discussion of a dataset on the effect of context on the speechreading of spoken sentences. *Journal of the Academy of Rehabilitative Audiology, 33*, 53-61.

Hutchinson, T.P., & Lai, P.W. (1981). Statistical aspects of

injury severity, Part III: Making allowance for differences in the assessment of level of trauma. *Transportation Science, 15*, 297-305.

Kidwell, S.M., Rothfus, T.A., & Best, M.M.R. (2001). Sensitivity of taphonomic signatures to sample size, sieve size, damage scoring system, and target taxa. *Palaios, 16*, 26-52.

Mietus, J.E., Peng, C.-K., Henry, I., Goldsmith, R.L., & Goldberger, A.L. (2002). The pNNx files: Re-examining a widely used heart rate variability measure. *Heart, 88*, 378-380.

Shallice, T., Rumiati, R.I., & Zadini, A. (2000). The selective impairment of the phonological output buffer. *Cognitive Neuropsychology, 17*, 517-546.

Swets, J., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Wang, H. (1997). Treatment of "don't know" responses in contingent valuation surveys: A random valuation model. *Journal of Environmental Economics and Management, 32*, 219-232.