ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

WILEY

Predictability and Prediction
Author(s): A. S. C. Ehrenberg and J. A. Bound
Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 156, No. 2 (1993), pp. 167-206
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2982727
Accessed: 10-10-2019 04:19 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms

JSTOR

Royal Statistical Society, *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series A (Statistics in Society)*

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

J. R. Statist. Soc. A (1993)
156, *Part* 2, *pp.* 167–206

Predictability and Prediction

By A. S. C. EHRENBERG† and J. A. BOUND

South Bank University, London, UK, and Stern School of Business, New York, USA

London Business School, UK

[*Read before* The Royal Statistical Society *on Wednesday, November 18th, 1992, the President,* Professor T. M. F. Smith, *in the Chair*]

SUMMARY

A result can be regarded as routinely predictable when it has recurred consistently under a known range of different conditions. This depends on the previous analysis of many sets of data, drawn from different populations. There is no such basis of extensive experience when a prediction is derived from the analysis of only a single set of data. Yet that is what is mainly discussed in our statistical texts. The paper discusses the design and analysis of studies aimed at achieving routinely predictable results. It uses two running case history examples.

Keywords: BETWEEN-GROUP ANALYSIS; CHOICE OF MODEL; CONDITIONS OF OBSERVATION; INDUCTION; LEAST SQUARES; MANY SETS OF DATA; PARSIMONY; PHYSICS; SOCIAL SCIENCE

1. INTRODUCTION

Many phenomena are routinely predictable, meaning that we can expect the outcome to recur with near certainty. Exceptions have to be similarly predictable as exceptions, e.g. that Boyle's law *never* holds for high pressures or when there is a leak in the apparatus. Routine predictability seems to arise simply from experience—even in everyday life—namely when the same outcome has occurred before for many sets of data (MSOD) observed under very different conditions.

Yet in statistics we traditionally regard prediction as difficult. We provide special analytical techniques which are supposed to give the 'best' prediction, although paradoxically they are expected to do so instantly, from just a single set of data (SSOD). But no justification is given for this in the literature. Nor is there a single quotable success story—a statistical prediction which has worked—when there should be hundreds or thousands.

Among the reasons for this extreme divergence between theory and practice is that there are many different kinds of prediction (e.g. Ziman (1991)). As we see it, some are interpolations into what is already known, others are more like extrapolations:

(a) *interpolative predictions*
  (i) inductive—inference to another population, or a sample therefrom, within the range of populations or conditions already covered;
  (ii) deductive—inferences about random samples from the same population;
  (iii) theoretical—routine deductions of a well-established finding (e.g. Neptune's position at 5 a.m. tomorrow morning);

† *Address for correspondence*: South Bank University Business School, 103 Borough Road, London, SE1 0AA, UK.

© 1993 Royal Statistical Society 0035–9238/93/156167

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

(iv) individual—a doctor's prognosis for a self-selected patient, based on experience;
(v) 'nowcasting'—predicting y from x for the data which the model has actually been fitted to (as also in saying 'It's raining *now*');

(b) *extrapolatory predictions*
(vi) forecasting—assertions about the future (e.g. next week's weather), when even the past patterns are still unclear;
(vii) discoveries—theory pointing to something that has not yet been observed (e.g. the very existence of the planet Neptune in 1846);
(viii) 'what-if?'—varying the input assumptions, as in using a spreadsheet,

and so on, up to and including soothsaying. These different predictive purposes are quite often confused.

This paper tries to clarify predictions of the first, the inductive kind, implying routinely predictable results. Section 2 first introduces the possibility of having such results at all. Section 3 then gives a brief overview of the relevant statistical literature. Sections 4 and 5 discuss the design and the analysis of predictive studies based on MSOD. Finally, Section 6 briefly considers the wider context.

To illustrate routine predictability we shall refer at times to the origin of Boyle's law in physics 330 years ago. This says that the pressure P of a body of gas varies inversely with its volume V under a great variety of conditions, i.e. that $PV \simeq C$, a constant for that body of gas and for a given temperature T.

We shall also use the 'double-jeopardy' (DJ) effect from social science, mainly marketing. This concerns people's choices between similar items such as different brands of a grocery product, different car makes or different politicians. The smaller of two brands, say, is not only bought by fewer people in a given time period (its penetration b), but this smaller number of customers also generally buy it less often (their average frequency of purchase w). A theoretical relationship between w and b is $w(1-b) \simeq w_0$, a constant for that product category and a given length of time period t.

What is being predicted in each case is not a particular value of an individual variable such as P or V, b or w, but the relationships between them. Success seems to depend more on the right choice of model than on parameter estimation. As Cobb (1991) has said 'Statisticians now put more effort into the complex process of choosing suitable models, less effort into doing those things—simpler by comparison—which take the choice of model as given'. Our paper is part of any move in that direction.

2. THE MAIN ISSUE

'Prediction' in this paper means asserting that the result in question will hold for some other, different, data (e.g. for a sample from another population). 'Predictability' means that it will hold routinely, within the relevant range of conditions.

The critical issue is the choice of model form. A small segment of a large circle or of a hyperbola can give as close a fit as any straight line. But which model will also hold for *other* sets of data? We use our two case histories to introduce three points here:

(a) that a model may fit an SSOD very well indeed but not hold at all for all or any of the many other sets of data that exist (Section 2.1);

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

(b) that another model can none-the-less hold for the MSOD in question (Section 2.2);
(c) that the difference comes from whether a model has been chosen which holds for more than one isolated SSOD in the first place (Section 2.3).

A fuller discussion is given in Sections 4 and 5.

2.1. *Starting from Scratch: Single Set of Data versus Many Sets of Data*

To illustrate, suppose that we are faced with the data in Table 1 for b, the penetration of a brand, i.e. the proportion of the population which bought the brand at least once in the analysis period, and w, the average frequency of purchase per buyer in the period, for each of the five leading brands of UK breakfast cereals in quarter 1 of 1968 (Aske Research, 1969).

The linear equation $w = 4.9b + 1.9$ fits well (see Section 5.3 for the method of fitting used), with deviations $(w - 4.9b - 1.9)$ which are irregular and average at 0.2 on the w-scale. The r^2-value is 0.90. But will the equation also hold for other data—can it become predictable?

We can first of all test this for the same five brands in the next quarter of 1968. Here the same equation holds within almost the same limits of scatter (± 0.3) and an r^2-value of 0.94. But this is not much of a test since the conditions in quarter II as well as the readings themselves were almost identical with those in quarter I anyway, a near 'hold-out' sample.

At the other extreme, we can test the model's predictability for something quite different, say for the five leading brands or brand groupings of cooking oil in Japan almost 15 years later as in Table 2 (Barnard *et al.*, 1993), i.e. for a different product, country, point in time, observers and rather different numbers also (especially for b).

The smaller brands again have fewer buyers (b) who buy them less often (w), the correlation being 0.98. The earlier correlation between b and w in Table 1 was therefore not just due to some fortuitous juxtaposition of the five UK cereal brands in 1968, but might generalize and become routinely predictable under radically different conditions (e.g. as the general DJ type of effect that was mentioned in Section 1).

TABLE 1
Breakfast cereals in the UK: quarter I, London, 1968†

	Proportion buying b	*Average no. of purchases per buyer w*	
Leading brands‡	*Observed*	*Observed*	***Fitted***
Corn Flakes	0.42	3.7	**4.0**
Weetabix	0.27	3.3	**3.2**
Shredded Wheat	0.16	2.9	**2.7**
Rice Krispies	0.17	2.5	**2.7**
Sugar Puffs	0.11	2.5	**2.4**
Average	0.23	3.0	**3.0**

†Observed b and w and the **fitted** equation $w = 4.9b + 1.9$.
‡In market share order.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

TABLE 2
Cooking oil in Japan: annual, Tokyo, 1983†

	Proportion buying b	*Average no. of purchases per buyer w*	
Leading brands‡	*Observed*	*Obser-ved*	***Predic-ted***
Nissin	0.70	4.7	**5.3**
'Other brands'	0.52	3.6	**4.4**
Aji-No-Moto	0.52	3.2	**4.4**
Ho Nen	0.41	3.0	**3.9**
Showa	0.23	2.1	**3.0**
Average	0.48	3.3	**4.2**

†Observed b and w and **predicted** $w = 4.9b + 1.9$.
‡In market share order.

Quantitatively the earlier UK equation $w = 4.9b + 1.9 \pm 0.2$, however, does not hold at all in Table 2. The predictions are all too high, mostly by about 0.9, and these deviations are highly significant with a sample size $n \simeq 1000$ (see Ehrenberg (1972), Table B4, for the standard error technicalities).

The equation cannot therefore be used for prediction. But should we perhaps have tested it in a less extreme way? Table 3 does so for the 1968 UK cereals brands again, but for b and w in a month or in the whole year. The DJ effect again recurs, but the numerical predictions of w are even worse than for Japan, all again about 1 unit too high in the monthly data, and up to some 5 units too *low* in the year.

One possibility is that there is no generalizable and hence routinely predictable two-variable relationship for w and b. Another is that $w = 4.9b + 1.9$ was the wrong form of model, i.e. that there can be *another* model which holds without systematic biases for *all* these data, as we now illustrate.

TABLE 3
Breakfast cereals in the average month and the year†

	Month			***Year***		
	b	*w*		*b*	*w*	
Leading brands	*Observed*	*Obser-ved*	***Predic-ted***	*Observed*	*Obser-ved*	***Predic-ted***
Cornflakes	0.28	1.8	**3.3**	0.62	10.1	**4.9**
Weetabix	0.17	1.7	**2.7**	0.41	8.5	**3.9**
Shredded Wheat	0.10	1.6	**2.4**	0.29	6.4	**3.3**
Rice Krispies	0.09	1.5	**2.3**	0.32	5.0	**3.5**
Sugar Puffs	0.06	1.5	**2.2**	0.23	4.8	**3.0**
Average	0.14	1.6	**2.6**	0.37	7.0	**3.7**

†Observed b and w and **predicted** $w = 4.9b + 1.9$.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

2.2. *Starting with Routinely Predictive Model*

Suppose that we already knew from the literature that the model

$$w_A(1-b_A) \simeq w_B(1-b_B) \simeq w_0,$$

held for any two competitive brands A and B (e.g. see Ehrenberg (1972) and Ehrenberg *et al.* (1990)). This says that the average frequency of purchase w should be smaller for a smaller penetration b (i.e. DJ), but not *linearly* so, i.e. that w varies as $1/(1-b)$. The w_0 is a calibration coefficient which can be estimated as the average value of $w(1-b)$, as we discuss further in Section 5.4.

We are told that the model has already held in a wide range of product categories, countries, analysis periods, points in time, etc. We can therefore successfully predict that it will also hold for the data in Tables 1–3, as is shown in Table 4. The fit is within small, mostly irregular looking deviations, averaging at 0.4, but somewhat heteroscedastic, an r^2-value of 0.95 overall. (The slight tendencies here to overpredict w for the market leaders and to underpredict for the second biggest brands do not in fact generalize to other data.)

The contrast with the previous analysis is that we now have a model $w(1-b) \simeq w_0$ which holds for all these different data sets, instead of the relationship $w \simeq 4.9b + 1.9$ which held rather more closely for the quarterly data as in Table 1, but not at all for the others.

2.2.1. *Boyle's example*

Few cases of starting with something like the right predictive model are documented in the statistical literature. As a different illustration, Boyle (1662) in his famous experiment already knew what he was doing, i.e. he had the prior hypothesis that pressure P and volume V would vary 'in reciprocal proportion'. (He probably had certain kinds of prior evidence even—Cohen (1964).)

TABLE 4

Test of the predictions for cereals and cooking oil†

	Breakfast cereals						*Cooking oil*		*Average O – P*	
	Monthly		*Quarterly*		*Annual*		*Annual*		*With sign*	*Without sign*
Leading brands	*O*	***P***	*O*	***P***	*O*	***P***	*O*	***P***	*O – P*	*|O – P|*
1st	1.8	**1.9**	3.7	**4.0**	10.1	**10.8**	4.7	**5.3**	−0.4	0.4
2nd	1.7	**1.7**	3.3	**3.2**	8.5	**6.9**	3.6	**3.3**	0.5	0.5
3rd	1.6	**1.6**	2.9	**2.7**	6.4	**5.8**	3.2	**3.3**	0.2	0.2
4th	1.5	**1.5**	2.5	**2.8**	5.0	**6.0**	3.0	**2.7**	−0.2	0.4
5th	1.5	**1.5**	2.5	**2.6**	4.8	**5.3**	2.1	**2.1**	−0.2	0.2
Average	(1.6)‡	**1.6**	(3.0)‡	**3.1**	(7.0)‡	**7.0**	(3.3)‡	**3.3**	0.0	0.3

†*O*, observed; *P*, **predicted**.

‡ Effectively used in estimating w_0.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

TABLE 5
Boyle's original experiment for $PV = C$†

V	*P*		*V*	*P*	
Observed	*Observed*	***Predicted***	*Observed*	*Observed*	***Predicted***
12.0	(29.1)	**(29.1)**‡			
11.5	30.6	**30.4**	5.8	61.3	**60.2**
11.0	31.9	**31.7**	5.5	64.1	**63.5**
10.5	33.5	**33.3**	5.3	67.1	**65.9**
10.0	35.3	**34.9**	5.0	70.7	**69.8**
9.5	37.0	**36.8**	4.8	74.1	**72.8**
9.0	39.3	**38.8**	4.5	77.9	**77.6**
8.5	41.6	**41.1**	4.3	82.8	**81.2**
8.0	44.2	**43.7**	4.0	87.9	**87.3**
7.5	47.1	**46.6**	3.8	93.1	**91.9**
7.0	50.3	**49.9**	3.5	100.4	**99.8**
6.5	54.3	**53.7**	3.3	107.8	**105.8**
6.0	58.8	**58.2**	3.0	117.6	**116.4**

†Original readings in eighths of an inch, here rounded. Predicted $P = 349.2/V$.
‡Used in fitting.

Table 5 gives Boyle's observed P and predictions C/V for 25 values of V of a given amount of air. He determined the value of C as 349 from his first pair of readings ($V = 12$, $P = 29\frac{1}{8}$ in, i.e. atmospheric pressure).

Boyle's results have since been played down by some as 'but loosely demonstrated' (see Geary (1943)). Yet his readings represent an r^2-value of 0.9998 in our modern jargon. He himself was aware of the issue of errors but appeared relaxed:

> 'Now although we deny not, but that in our table some particulars do not so exactly answer to what our formerly mentioned hypothesis might perchance invite the reader to expect; yet the variations are not so considerable, but that they may probably enough be ascribed to some such want of exactness as in such nice experiments is scarce avoidable'.

2.3. *Need for Analysing Many Different Sets of Data*

Boyle, however, denied that his result $PV \simeq C$ was as yet predictable, despite its good fit:

> 'But for all that, till further trial hath more clearly informed me, I shall not venture to determine whether or not the intimated theory will hold universally and precisely either in condensation of air, or rarefaction'.

What mattered was how far his 'law' would be found to hold in further trials for air (and ultimately also for other kinds and mixtures of gases), for different amounts of gas, for pressure going up or coming down (for which the highly inventive Boyle devised quite a different initial experiment), for different apparatus, at different temperatures, for different (experienced) observers, in different locations, at different points in time, and so on.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

It had also to be established that *failures* were predictable, as cases where $PV \simeq C$ *never* holds, e.g. for high pressures (which Boyle wanted to test but technically could not, and about which his conjectures were wrong), for changing temperatures (which he tried to test at least minimally), when liquefaction occurs (e.g. of some water vapour), or osmosis or the like (which he took care to avoid), or a leak (which he also sought to avoid). It was not merely the rule that had to be predictable, but also the exceptions.

The predictability of the DJ model $w(1-b) \simeq w_0$ rests similarly on the range of conditions under which it (and its systematic exceptions) have been empirically established. This is summarized in Table 6, and is still increasing. What matters is not only that DJ holds for particular products—e.g. petrol as well as breakfast cereals—but how the conditions differ in other respects too: food for people or for cars, differentiated brands (anyone can tell a Corn Flake from a Shredded Wheat) *versus* ones which are not (even the motor-car cannot usually tell the difference between Shell and Esso), brands sold side by side or in solus-site distribution, sold in packets or 'loose', at very different prices (£1 or £10 or so, or even £10 000 for a motor-car), and so on. Yet the DJ pattern still holds and has therefore become predictable within this even wider range of different conditions.

TABLE 6
Conditions under which DJ is known to occur

Food and drink
Biscuits, butter, canned vegetables, cat and dog foods, cigarettes, coffee, confectionery, convenience foods, cooking fats and oil, flour, food drinks, frozen foods, instant potatoes, jams and jellies, margarine, pet foods, processed cheese, sausages, soft drinks, soup, spreads, take-home beer

Cleaners and personal care products
Cosmetics, deodorants, detergents, disinfectants, disposable nappies, feminine protection, hair-sprays, household soap, household cleaners, paper tissues, polishes, shampoos, toilet-paper, toilet-soap, tooth-paste, washing-up liquids

Automotive, health, information, entertainment and financial
Aviation fuel, motor-cars, motor oil, petrol;
over-the-counter medicines, pharmaceutical prescriptions;
newspapers, politicians, television programmes;
financial services

Distribution channels
Chains, individual stores, brands within chains, television channels

Time, place and measures
1950 to 1991;
Britain, continental Europe, USA, Japan;
demographic subgroups;
analysis periods from 1 week to 2 years;
recorded and reported buying

Exceptions or partial exceptions
Restricted distribution, private labels;
some submarkets;
special funding (e.g. luxury cars, bible stations, bribes)

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

3. THE STATISTICAL LITERATURE

In this section we give an overview of what the statistical literature seems to say on issues of routine predictability. This can be brief because little appears to have been said. We cover comments on

(a) the need for experience (Section 3.1),
(b) how prediction differs from statistical inference (Section 3.2) and
(c) the statistical use of prior knowledge (e.g. Bayes) (Section 3.3).

3.1. *Experience and Varied Conditions*

In 1989 Mrs Thatcher's then government in its new national curricula for UK schools prescribed that 'At levels one and two [infant school] children will learn to make predictions based on experience' (News & Notes, 1989). Such references to 'experience' are, however, rare in our statistical literature (e.g. texts for adults), though even more emphatic when they *do* occur (Fisher (1926), our italics):

> '*Most important of all*, the conclusions drawn . . . will be given, by the variation of non-essential conditions, a very much wider basis than could be obtained . . . without extensive repetitions of the experiment'.

Tukey (1989), in quoting Fisher, rightly added (our italics again):

> 'This [need for extensive repetitions] is unhappy for the investigator who would like to settle things once and for all. But [it] is consistent with the best accounts we have of the scientific method, which emphasise repetition—*preferably under varied circumstances*.'

Unfortunately that appears to be about all that Fisher and Tukey said about the need for experience. (Fisher emphasized instead the reduction of sampling errors—see also for instance Smith (1991).) Yet the idea of varying such 'non-essential' conditions of observation is commonplace (as Tukey says) in science.

3.2. *Prediction versus Statistical Inference*

Modern statistics largely focuses on 'drawing inferences . . . from the sample to the population' (Fisher, 1935). This deductive form of inference, however, arises mainly with small samples and almost disappears with large samples, and also cumulatively with MSOD.

How such inference relates to prediction in any broader sense is almost never discussed. But Lindley (1947) and Fuller (1987) say that regression predicts the value of y for another reading of x sampled (in an apparently unspecified way!) *from the same population*. This is probably also what many other statisticians believe, though they seldom say it explicitly.

However, that is not prediction in our sense here. Nor is it, we believe, prediction for the person or scientist in the street. Inductive inference (type (i) in Section 1) is usually meant to be to *new* data, i.e. to a different population (e.g. Deming (1992)). It has to be what Popper calls falsifiable, like that $w \simeq 4.9b + 1.9$ does *not* hold for the other data in Section 2.1. Aitchison and Dunsmore (1975) have perhaps uniquely stressed that the data to which a prediction is to be made needs to be independent of the data on which the prediction is based (not 'nowcasting'). This led them explicitly to exclude standard time series analyses and least squares regression as not giving a prediction, i.e. an assertion about different, independent, data.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Advanced statistical texts tend to distinguish regressions for prediction from structural equations for scientific laws (e.g. Kendall and Stuart (1979)). But they do not give usable forms of structural analysis nor prove why there should be any such distinction anyway (*scientists* use their laws for prediction!).

3.3. *Statistical Prior Knowledge and Superpopulations*

Bayesians and others have tried to move beyond analysing an SSOD by bringing in prior knowledge (or judgment). This seems highly desirable, if not essential, but not the Bayesians' probabilistic apparatus (despite a long philosophic tradition). The relationships $PV \simeq C$ or $w(1-b) \simeq w_0$ have each held for a wide range of different previous conditions and will do so again. If not, something totally new will have been discovered. This is not probabilistic, as far as we can see.

Other moves to use prior information include 'shrinkage' (e.g. Copas (1983), Zellner and Hong (1989), Berger (1985) and Miller (1990)), various developments in econometric modelling (see Judge *et al.* (1985) and Kennedy (1991)) and constrained applications of generalized linear models for dealing with structured data (e.g. McCullagh and Nelder (1989) and Hastie and Tibshirani (1990)). But the focus is on estimating distinct parameters for each separate data set and then comparing the results. If such approaches worked predictively there should be many successful examples. Granger's 'co-integration' approach (e.g. Cuthbertson *et al.* (1992)), however, has some similarities with the MSOD approach here. For two time series $x(t)$ and $y(t)$ it looks for local means $\bar{x}$ and $\bar{y}$ at two different times t_1 and t_2 where the means are very different, so that the $\bar{x}/\bar{y}$ correlation is very high and most problems with 'best fit' regression disappear. But the data are still treated by ordinary least squares as an SSOD.

So-called 'meta-analysis' (e.g. Hedges and Olkin (1985), Christie (1990), Leftwich (1990) and Wachter and Straf (1990)) stands for the idea of combining information from all the relevant data sets (Fisher (1935) stressed the *all*). This should be virtually commonplace. But meta-analysis mostly emphasizes the reduction of sampling errors (usually for a borderline result), rather than looking for a single model. As and when meta-analysis focuses on successfully finding generalizable relationships across MSOD (i.e. looking for 'significant sameness'—Nelder (1986)), it will virtually coincide with our approach in this paper, which we in turn regard as largely typical of normal science.

There is also talk of superpopulations: 'The entire result of an extensive experiment may be regarded as but one of a possible population of such experiments' (Fisher, 1925). But how is a superpopulation to be defined, enumerated and sampled? Eisenhart (1946) asked whether a new sample would be drawn from the same population as before, but he gave no answer: he merely hoped that 'raising this query would not only reduce the reader's headaches . . . but also lead him to the correct decisions'. It seems doubtful whether that has yet happened (e.g. Robinson (1991)).

One of the referees has said (our italics) that

> 'much of current statistical methodology, where it has acknowledged prediction as a feature, has considered predictions of future values from the same population of study—or more accurately [sic!] *a future population of the same kind*'.

But we can hardly base scientific predictions merely on some undefined subjective judgment or 'leap of faith' (Draper *et al.*, 1993) to decide whether different populations

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

are 'of the same kind'. At best it merely seems a circular argument, that we can predict that the results will be the same only when we have already somehow felt able to assume that the populations are of the same kind anyway.

4. DESIGN OF PREDICTIVE STUDIES

In this paper we are saying that an observed result such as $PV \simeq C$ or $w(1-b) \simeq w_0$ becomes routinely predictable only when it is based on extensive experience. The crux is not whether different populations are somehow 'of the same kind', but whether the *result* in the past has turned out to be the same. We therefore consider in this section the design of studies that can make this happen.

4.1. *A Wide Range of Conditions*

Predictability in Section 2 rested on the wide range of conditions covered, in line with the Fisher or Tukey asides 'Most important of all . . . are extensive repetitions of the experiment . . . under varied circumstances'.

Replications or repetitions can never be strictly identical (although this is an ill-conceived popular ideal) since at least the time and/or place must have changed. In any case, the results would then also have to be identical and we would have learnt nothing. Instead of trying to sweep supposedly minor differences under the carpet, our view is that such variations in Fisher's 'inessential conditions' are altogether of the essence: the same result has held *despite* such and such differences in the conditions of observation.

The basic design precept has, we think, to be 'When in doubt, find out', i.e. vary the factor (or let it vary) and see whether it affects the outcome. To do this well requires care and effort. There can be no instant solutions, as Tukey (1989) warned.

Much depends on the 'nitty-gritty' detail. We could for example assess that *brand images* need not interfere with the predictability of *double jeopardy*, because DJ has held for breakfast cereals, where the whole family can see the brand and possibly be affected by its 'image', as well as for cooking oil, where they are usually not even aware of the brand.

Time is often regarded as a special complication: 'Prediction is difficult, especially when it concerns the future'. But we think not. Neither DJ nor Boyle's law has in the past been affected by time: $w(1-b) \simeq w_0$ for example held both in 1968 and in 1992. If it does not recur next year say, this cannot just be due to 'time' as such, but some much more specific new factor, such as a poison scare then, a new European Community directive outlawing DJ as being anticompetitive, or whatever, but no longer just time.

Replications can be thought of as of broadly two kinds (e.g. Lindsay and Ehrenberg (1992)):

(a) *close*—design factors are varied where we do not expect this to affect the result (if we turn out to be wrong, that would be highly informative); close replications tend to be relatively cheap and easy; they are useful for instance in checking whether a new result is repeatable at all;
(b) *differentiated*—the study is repeated under distinctly different conditions, say DJ in an underdeveloped economy or for buying by mail order; this is where one either radically extends or limits the range of predictability of the findings.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

But the distinction varies over time. What may initially be differentiated (or speculative)—such as in Sections 2.1 and 2.2 whether DJ would also hold for cooking oil in Japan—becomes a close replication once it has been found to have done so.

4.1.1. *Additional design factors*

Additional factors can and should always be varied, often at little extra cost, in studies which have been designed for some other specific purpose. If the variation makes no difference to the results, we have established or confirmed that without any loss in the accuracy or sample size for our main purpose. If, however, it leads to different results, the sooner we know that the better.

There is also much room for 'confounding' (unlike with causal studies) and 'partial' replications. Thus, in going from Table 1 to Table 2 earlier, the product (cereals to cooking oil), the country (UK to Japan), the time (1968 to 1983) and the analysis period (a quarter to a year) all changed together (as well as other unstated conditions). These factors were 'confounded'. But since $w(1-b) \simeq w_0$ still held it seems that *none* of the factors mattered, within the limits of approximation. Possible compensating effects will be increasingly ruled out by other, later replications (e.g. for breakfast cereals and/or detergents in Japan and/or Korea).

4.1.2. *Deviations and exceptions*

Deliberate replication is needed to pin down a deviation: first a close replication to see whether the deviation recurs at all. For example, in the quarter I London data of Table 1, Rice Krispies and Sugar Puffs had the same observed w of 2.5, despite marked differences in b (0.17 and 0.11). But in the quarter II data (not shown) the two ws differed appropriately. The quarter I discrepancy was therefore not related to the brands as such, but to some other momentary factor. In contrast, private label brands in the UK tend more consistently to have a somewhat higher average buying frequency w (although the causal explanation is different—Ellis (1989) and Ellis *et al.* (1993)).

4.2. *Observation, Experimentation and Control*

The conditions of observation in replicated studies differ in one of two ways:

(a) *observationally*, by the deliberate selection of naturally occurring conditions of observation (e.g. large *versus* small brands in Tables 1–3, cereals *versus* cooking oil, Japan *versus* the UK, 1983 *versus* 1968, and so on);
(b) *experimentally*, by the deliberate variation of a factor such as the amount of mercury one pours into a U-shaped tube which is sealed at one end; this may seem a peculiar thing for grown men to do, i.e. it is definitely 'experimental' and presumably done with a purpose.

Thus Boyle and his collaborator Robert Hooke

> 'continued this pouring in of quicksilver till the air in the shorter leg was reduced to take up but half the space it possessed (I say possessed, not filled) before; we cast our eyes upon the longer leg and observed, not without delight and satisfaction, that the quicksilver in that longer part of the tube was 29 inches higher than the other'.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Boyle next aimed at the *fourfold* increase of the pressure that was shown in Table 5. His famous experiment was therefore only a rather prosaic but very detailed replication (the initial excitement was over). It was done

> 'because an accurate experiment of this nature would be of great importance to the doctrine of the spring of the air, and has not yet been made (that I know) by any man'.

More of the data in the physical and biological sciences is, however, purely observational than is widely thought (e.g. *all* in astronomy), and in the social sciences more is deliberately experimental.

4.2.1. *Factors to control*

'Control' means varying some conditions of observation, and also keeping those the same which might or do affect the result, like the temperature for gas, or retail availability for consumer products. (In later replications we would let some of these factors vary, precisely to see that they *do* matter, and also *how*.)

But not *all* extraneous influences and error possibilities are usually controlled. Boyle had for instance managed, with some difficulty, to construct a crooked tube

> 'so tall, that we were fain to use it on a pair of stairs being two to make the observation together, the one to take notice at the bottom, how the quicksilver rose in the shorter cylinder, and the other person to pour in at the top of the longer, it being very hard and troublesome for one man alone to do both accurately'.

But Boyle could have poured in just *any* additional amounts of mercury into his glass tube and simply measured P and V each time, both as uncontrolled variables. Other design conditions could also have varied and some did, as when 'we were hindered from proceeding at that time by the casual breaking of the tube'. The crux lay in that Boyle had picked much the right variables and conditions in the first place, and then in his management of the experimental situation ('we took, I say, care').

'Experimentation' merely means deliberately and artificially varying a factor to see what will happen ('even God himself may be waiting to see'—Ziman (1991)). It need not involve control groups. These seem to be mainly called for in 'engineering' applications (e.g. agricultural or clinical trials) if what would happen without the treatment is not yet routinely predictable, or not sufficiently precise.

The role of the fully *randomized* experiment also seems more peripheral than most statistical texts imply. It can eliminate the systematic influence of otherwise uncontrolled factors and should speed up the isolation of causal effects. But randomization cannot be of the essence when building up a routinely predictable result from MSOD. Differences between different randomized experiments cannot be randomized away (e.g. Japan *versus* Britain, cooking oil *versus* cereals). Nor do we usually want to eliminate them, but rather to show that the same result held *despite* such differences.

Selecting or creating highly differentiated replications leads to very high correlations, as in our examples. None-the-less, even Boyle's studies were constrained not only by how far he knew what to control and how to control it, but also by his budget:

> 'If we had been furnished with a greater quantity of quicksilver and a very strong tube, we might have . . .'.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

5. ANALYSING THE DATA

Data analysis in the present context means mainly either testing or extending the predictability of an already given result. For example, after Boyle had established that volume halved when the pressure doubled ('not without delight and satisfaction'), came the much staider business of establishing under what much wider conditions, and in just what quantitative forms, this proportionality effect recurred. As Boyle himself recognized, the form of analysis required was merely to see whether the given relationship held again. This we illustrate briefly with the DJ model $w(1-b) \simeq w_0$ in Section 5.1. It is not a case of having to derive or estimate any *new* relationships.

The initial derivation of a new model is none-the-less also of interest. In Section 5.2 we therefore consider this in terms of two or more sets of data (MSOD). Success depends largely on having hit on the right *form* of the relationship, as we discuss in Section 5.3.

At no stage in analysing MSOD is there any call, as far as we can see, for the optimizing (e.g. best fit or maximum likelihood) principles of classical statistics. Instead, the basic criterion is whether the same relationship holds for all the different data sets. This arises mainly as a matter of parsimony, because a model which differs from instance to instance would be too difficult to validate, as we note in Section 5.4.

5.1. *Testing Further Data*

To extend the predictive scope of a given model we simply have to see whether or not it also holds for the new data. This means that the relationship should hold within much the same limits of scatter as before ('stochastic sameness'). It is not a question of having to judge 'how close is close' (Nelder, 1989) or whether the degree of approximation is somehow 'acceptable' (Draper *et al.*, 1993), but whether the scatter is itself also predictably about the same.

The residuals should ideally be irregular (so that they are easy to summarize as such). In practice, this criterion has to be relaxed to cope also with any consistent biases that have already been reported (e.g. for high pressures in the case of Boyle's law), outliers and heteroscedasticity (see also say Ziman (1991) about the 'fit' of theories in *physics*).

To check for instance whether DJ also holds for people's *store choice*, Table 7 gives the earliest store choice data analysed (Kau, 1981). The prediction was that w would decrease as $w_0/(1-b)$, with w_0 being the average of the observed $w(1-b)$, 2.6, for the five store groups.

There are three steps in the analysis of such a table (which should perhaps be formalized further), namely to note whether

(a) there is any overall bias (i.e. $w = 3.2$ observed and predicted here),
(b) the individual deviations appear irregular (e.g. the observed and predicted w for the top two chains broadly agree at 3.65 and 3.45, and for the bottom two at 2.85) and
(c) the size of the deviations (averaging at 0.4, a residual standard deviation of 0.5) is virtually in line with those that occurred before (e.g. in Table 4).

The correlation in Table 7 between b and w is only 0.71, because the w vary relatively little. But DJ for store choice has also been found for other products and in other countries, and with higher correlations if there was more variation in w (e.g. Kau

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

TABLE 7
Testing DJ predictions $w_0/(1-b)$ for store choice: buying instant coffee at leading chains, Lancashire, half-year, 1979

	Proportion buying b	*Average no. of purchases per buyer w*	
Store	*Observed*	*Observed*	***Predicted***
Coop	0.28	3.4	**3.6**
Kwiksave	0.22	3.9	**3.3**
Tesco	0.21	3.1	**3.3**
Asda	0.14	3.6	**3.0**
Fine Fare	0.05	2.1	**2.7**
Average chain	0.18	3.2	**3.2**

(1981), Wrigley and Dunn (1984), Uncles and Ehrenberg (1990) and Uncles and Hammond (1992)). The difference from classical bivariate analysis is that Kau and the later analysts did not have to fit a new model.

5.2. *Determining the Slope for Many Sets of Data by Between-group Analysis*

The original derivation of a relationship—e.g. how Boyle did it—is part of the history of science and not, we believe, of its substance. It arises infrequently (both scientists and engineers mostly use already established models). It differs in this and other respects from the instant best fit procedures of traditional statistics. In this section we discuss the fitting of a new straight line model to MSOD (possibly after transforming one or both variables).

Classical bivariate regression fits a line to an SSOD, A say, with a pair of means $(\bar{x}_A, \bar{y}_A)$. It does so by bringing in a 'best fit' criterion like least squares, applied to the uncontrolled scatter of the individual readings in A. But fitting a regression line to describe the conditional expectation $E(y|x)$ in a single bivariate distribution (*A* say) is not the same problem as describing a linear relationship between the means $\bar{x}$ and $\bar{y}$ in two or more different bivariate distributions, *A*, *B*, *C*, etc. The latter is quite simple (at least in the exactly linear case), but seems worth spelling out.

When fitting a line to two or more distinct sets of data, A, B, C, etc., the basic question is how the pairs of means $(\bar{x}_A, \bar{y}_A)$, $(\bar{x}_B, \bar{y}_B)$, $(\bar{x}_C, \bar{y}_C)$, etc. vary together. To the statistician this may be thought of as the bivariate equivalent to univariate analysis of variance (but with the emphasis on modelling the variation of the means rather than on only testing it for significance against the within-group scatter). The systematic covariation of the paired means determines the parameters. For just two pairs of means $(\bar{x}_A, \bar{y}_A)$ and $(\bar{x}_B, \bar{y}_B)$, the slope has to be $(\bar{y}_A - \bar{y}_B) / (\bar{x}_A - \bar{x}_B)$. This is the simplest instance of what has been called between-group analysis (BGA), which describes how the means $\bar{x}$ and $\bar{y}$ vary together between different groups of data.

With more than two data sets A, B, C, D and E, say, the slope coefficient could be determined from any two if all the paired means lay *exactly* on a straight line, or almost so as with Boyle's Table 5. In practice there will be some deviations. There are then many possible values of the slope even when each set of readings is the full

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

population data (or a large sample). This is not a statistical issue in the classical sense, because one is forcing a straight line on to *population* data known to be strictly non-linear (i.e. a matter of deliberate oversimplification, not one of merely smoothing out random or quasi-random sampling deviations from some underlying true model).

One possibility is to put a BGA line through the two extreme pairs of means ($\bar{x}_A$, $\bar{y}_A$) and ($\bar{x}_E$, $\bar{y}_E$) as in Fig. 1. If all the pairs of means follow an overall linear trend, this will tend to pick it up well. A slightly more robust BGA alternative is to estimate the slope from the averages of each of two or three such extreme pairs of means and the intercept coefficient from the overall means (e.g. Ehrenberg (1975), chapter 7, and Ehrenberg (1982a), chapter 13). This also deals with cases where there is no unique extreme (as for the w of Rice Krispies and Sugar Puffs in Table 1).

To illustrate such BGA for Table 1 earlier, we average the observed w and b for the two largest brands and those for the two smallest. This gives a slope of 4.9, i.e.

$$\frac{3.5-2.50}{0.345-0.140}=\frac{1.00}{0.205}=4.88, \text{ or } 4.9 \text{ rounded.}$$

Putting the line through the overall means ($w=3.0$, $b=0.23$) gives $w=4.9b+1.9$, as in Section 2.1. This BGA fitting process could no doubt be formalized, e.g. by some use of GLIM (McCullagh and Nelder, 1989).

Fitting by BGA is akin to the analysis proposed by Wald (1940) and Bartlett (1949) for two variables subject to error. But the design differs radically. Wald and Bartlett focused traditionally on an isolated SSOD (e.g. a bivariate normal distribution) and had to superimpose an arbitrary division to create their two or three subgroupings (with an insuperable lack of uniqueness even in an exactly linear case, as stressed by Wald). Fitting the actual line to the two or three pairs of subgroup means ($\bar{x}$, $\bar{y}$) is then altogether straightforward (i.e. technically 'trivial'). In contrast, in BGA for MSOD as here, we fit the slope to the different pairs of means ($\bar{x}$, $\bar{y}$) of distinct sets of data, i.e. groupings of the data that were determined by the observer before the event and not by the analyst *post hoc*. We can (or should) use observational or experimental designs which give markedly different means and therefore high r^2-values as noted in Section 4.4.

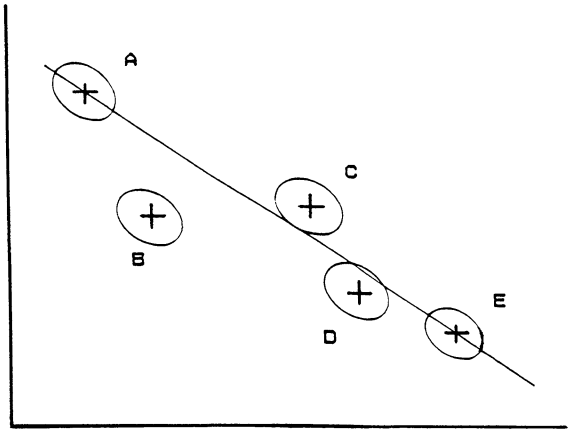

Fig. 1. Five different data sets A–E (BGA model fitted to the means $\bar{x}_A$, $\bar{y}_A$ and $\bar{x}_E$, $\bar{y}_E$)

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

The justification of any new BGA line is, however, not in the estimation technique, nor even in its fit (e.g. that $r^2 = 0.90$ for Table 1), but in whether it holds for all the different subsets of the data (the five brands here) without any marked systematic subpatterns or 'local biases'. This is the crucial criterion for BGA. Even then such a newly fitted equation is only a first 'working solution'. We discuss in Section 5.3 how it may need to be drastically adjusted later if it does not hold for *other* data, e.g. for other countries, products, or lengths of time periods, as in Tables 2 and 3.

In itself, *any* line fitted to Table 1 would be virtually trivial. Who now cares about the top five cereal brands in London in 1968 or about Boyle's 25 pairs of readings of P and V in 1662? As Boyle himself stressed, predictability depends on whether any resulting relationship has also held for a much wider range of conditions.

5.2.1. *Individual readings*

The variability of the individual readings within each data set is central to traditional best fit procedures (see also Section 6.2). But with BGA and MSOD this within-group variability is only a minor part of what is being predicted. It is often not even reported explicitly, let alone used when fitting the model. The individual readings to be analysed are mainly just summarized by their means, e.g. the values of b and w for each brand.

The within-group scatter should, however, also be described. Thus the DJ data in Tables 1–3 and 7 was for samples of n households (with n about 1000 or more). Within each sample the distribution of individual households' purchase rates about the mean w was always highly skew, but this is itself routinely predictable by a two-parameter negative binomial distribution with w and b as the two inputs, as shown in Table 8 (Ehrenberg, 1959, 1972; Kendall, 1961).

In contrast, Boyle's experimental results in Table 5 were for 25 distinct 'populations' which consisted of only a single reading each ($n = 1$). He did not generate samples of repeated readings even though he recognized the difficulty of 'making a just estimate of the true place of the protruberant mercury's surface'. He already knew from previous experience (i.e. could routinely predict) that the inherent variability of his material and his measurement errors would both be small, and no doubt more or less 'normally' distributed, as was verified by his close degree of fit in Table 5. (It is not always so in physics.)

More generally, the 'direction' of any within-group scatter could differ from that of the relationship between the group means, e.g. there could be *positive* correlations within each data set in Fig. 1 but still a *negative* slope for the overall relationship

TABLE 8

Frequency distribution of instant coffee purchases at Tesco: Lancashire, half-year, 1979†

Proportion buying b	*Number of purchases*							*Average* w
		1	*2*	*3*	*4*	*5*	*6+*	
0.21‡	Observed (%)	38	19	14	9	7	12	3.1‡
	Predicted (%)	**42**	**20**	**11**	**8**	**5**	**14**	

†Percentage of buyers buying instant coffee 1, 2, 3, etc. times in the half-year: observed and negative binomial distribution (Kau, 1981).

‡Used in fitting the negative binomial distribution (from Table 7).

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

between the means $\bar{x}$ and $\bar{y}$. This would again be a part of what we would have to predict for further such data and ultimately to explain.

There are also certain 'engineering' applications of law-like relationships which turn on an individual reading. A doctor's prognosis for an individual patient was noted in Section 1. In this paper our concern is, however, with the predictability of the relationship between aggregates (i.e. means), such as $w(1-b) \simeq w_0$ or $PV \simeq C$.

5.3. *Finding the Functional Form*

Finding a routinely predictable model depends on finding a suitable functional form. Thus the closely fitting linear equation $w = 4.9b + 1.9$ in Table 1 did not fit the other data in Section 2.1 at all. But the prior non-linear equation $w(1-b) \simeq w_0$ did. We now consider how such a model could be derived initially. This seems to involve insight and subject-matter knowledge, rather than formal statistical techniques. It can be helped by data analysis, or by theory, or both. *The rigour comes from testing any candidate model against yet further MSOD.*

5.3.1. *Data analytic derivation of* $w(1-b) = w_0$

For a start, we could fit a BGA equation for each set of five brands in Tables 1–3 (as was illustrated earlier for cereals in quarter I). This would give

Cereals, annual	(Table 3):	$w = 18.3b + 0.1$	$(r^2 = 0.85)$,
Cereals, quarter I	(Table 1):	$w = 4.9b + 1.9$	$(r^2 = 0.89)$,
Cereals, monthly	(Table 3):	$w = 1.7b + 1.4$	$(r^2 = 0.88)$,
Cooking oil, annual	(Table 2):	$w = 5.5b + 0.7$	$(r^2 = 0.95)$.

The fit of each equation is close ($r^2 \simeq 0.9$), but the equations differ greatly (and significantly) from each other. None therefore offers any predictability to other, different data. Which equation would we use?

The equations may differ because of excluded variables (e.g. explicit variables like the different products and countries, or implicit variables like price or advertising). Alternatively we may simply be using an inappropriate functional form. Indeed, although the data in each table are close to linear, looking at all the data in the three tables together shows up two non-linear subpatterns which we have to take into account in choosing a model.

(a) The average level of the ws differs far more between the four data sets than does that of the bs, e.g. in Table 3 from average ws of 7 annually down to 1.6 monthly, compared with average bs of 0.37 down to 0.14. Should we therefore perhaps model the *relative* variation in w from brand to brand, possibly by some index such as w/w_0? Here w_0 could be a suitably standardized value of w, such as, by some insight, the estimated purchase frequency w for each table when $b = 0$ (extrapolated by eye at this stage as about 1.6, 2.3, 4.1 and 1.4).

(b) The w vary less within each data set when the b are smaller. For example, w varies between the brands in Table 3 by about 5 for a difference in b of 0.4 annually, but by only 0.3 for a difference in b of 0.2 monthly. This suggests that we either introduce a third variable t for the length of the analysis period, or, more parsimoniously, a non-linear transformation of b such as, by some insight again, $1/(1-b)$. This cleverly varies less with b when the bs are smaller than when they are larger.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Fitting straight lines by BGA to the two transformed variables w/w_0 and $1/(1-b)$ gives

Cereals, annual:	$w/w_0 = 1.4/(1-b) - 0.8$	$(r^2 = 0.79)$,
Cereals, quarterly:	$w/w_0 = 1.1/(1-b) - 0.2$	$(r^2 = 0.87)$,
Cereals, monthly:	$w/w_0 = 0.8/(1-b) + 0.2$	$(r^2 = 0.90)$,
Cooking oil, annual:	$w/w_0 = 0.8/(1-b) + 0.4$	$(r^2 = 0.94)$.

These are much more alike in their slopes. A single equation might therefore fit all these data with little or no bias and hence be potentially predictable for yet other data. Since the four intercept coefficients tend to vary inversely with the fitted slopes, lines with a common intercept would also have more nearly equal slopes. Indeed, with an intercept of *0*, the result is four virtually identical equations:

Cereals, annual:	$w/w_0 = 1.00/(1-b)$	$(r^2 = 0.79)$,
Cereals, quarterly:	$w/w_0 = 0.98/(1-b)$	$(r^2 = 0.86)$,
Cereals, monthly:	$w/w_0 = 0.99/(1-b)$	$(r^2 = 0.90)$,
Cooking oil, annual:	$w/w_0 = 0.99/(1-b)$	$(r^2 = 0.94)$.

The slopes can be rounded to 1, giving the single model $w/w_0 = 1/(1-b)$, or $w(1-b) = w_0$.

The fit of this model was shown in Table 4 to be within ± 0.4 on average (an overall $r^2 = 0.94$) and with little (or ultimately no) biases. We therefore now have one model which holds for all the data, rather than a different model for each table which holds somewhat more closely still for that particular data but not at all for any other.

5.3.2. *Derivation of $w(1-b) = w_0$ by theory*

In practice, the model was initially derived by theory rather than by data analysis. Thus Goodhardt and Chatfield posed two independence assumptions several years ago, both of which were already known to be approximately true (e.g. see Ehrenberg (1972), section 11.5):

(a) that the buying of different brands, say A, B, C, etc. would be independent;
(b) that buyers of brand A would buy the whole product category (*any* brand) at the same average rate as would buyers of brands B or C, etc.

The relationship $w(1-b) = w_0$ then follows by simple algebra, as follows.

Consider for simplicity just three brands A, B and C. Brand A's buyers' average rate of buying the product category is then $(b_A w_A + b_{BA} w_{B.A} + b_{CA} w_{C.A})/b_A$, where b_{BA} is the proportion of the population who buy both brands B and A in the analysis period and $w_{B.A}$ is their average frequency of buying B, given that they also bought A. On assumption (a), we can write $b_{BA} = b_B b_A$ and $w_{B.A} = w_B$. Hence A's buyers' rate of buying the product category simplifies to $w_A + b_B w_B + b_C w_C$, and similarly we have $w_B + b_A w_A + b_C w_C$ for B's buyers' rate of buying the product category. On assumption (b), these two rates should be equal. Cancelling the common term $b_C w_C$ therefore gives $w_A + b_B w_B = w_B + b_A w_A$. Collecting terms in A and B then gives $w_A(1-b_A) = w_B(1-b_B)$, or $w(1-b) =$ a constant. This form of DJ therefore has to happen if assumptions (a) and (b) are true, as they approximately are.

5.4. *Few Variables and Fewer Parameters*

Both models in this paper involve only two 'essential' variables and even fewer parameters. This seems typical of many of the law-like relationships of science in their

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

more basic forms (e.g. Ohm's law, Newton's laws, $E=mc^2$, etc.). It is not so much that few is beautiful (although it is), but that validating a many-variable model from scratch would, we think, simply be too demanding. (No routinely predictable multiple-regression models seem to exist in the literature.)

With controlled designs (see Section 4.4), we can fit such simple (or oversimplified) models with very high r^2 but very few variables (if a generalizable relationship exists), leaving exceptions and outliers to be described as exceptions. Extra variables to boost a low r^2 are not needed.

Other variables may, however, be added later to account for consistent subpatterns. Validating such second-order factors may by then be feasible. For Boyle's law at high pressures we have for example to allow for the mutual attraction of the gas molecules (traditionally represented by A) and for their volume B, as in the 'van der Waals correction' $(P+A/V^2)/(V-B)\simeq C$.

5.4.1. *Fewer parameters*

The two case history models have even fewer parameters than variables. Expressing them in log-linear form brings out that both slope coefficients are absolute constants, at -1:

$$\begin{array}{ll} \textit{Boyle} & \log P = -\log V + \log C; \\ \textit{DJ} & \log w = -\log(1-b) + \log w_0. \end{array}$$

Having few if any numerical parameters seems again not just a case of parsimony for beauty's sake (although it is *very* elegant). It is that establishing the generalizability of a new multiparameter model across MSOD would again be difficult, if not usually impossible. It may be a form of natural selection where only the simplest models survive.

The slope coefficients of the relationships (β in general, but -1 in the two cases here) are the core issue in parameter estimation. The parameters C and w_0 merely reflect the average values of P or w in each array of data (i.e. they are 'calibration coefficients'). They vary for different gases, amounts of gas and temperatures T, or for different products (e.g. cereals *versus* cooking oil), countries and lengths of analysis period t. Given the slopes, the values of C and w_0 then typically just fall out when we put the line through the overall means of the data in question. (In classical statistical terms we still need to use 1 degree of freedom for each data set to estimate C or w_0, but none here for the more critical slope coefficients.)

Boyle, however, determined his C at 349 just from his first pair of readings (for P at atmospheric pressure). With $w(1-b)\simeq w_0$ we could similarly estimate w_0 from just one brand, but more robustly we would use the average value of $w(1-b)$ across the itemized brands or, for still greater statistical robustness, a *weighted* average such as $\Sigma\, bw(1-b)/\Sigma\, b$. (The equivalent for Boyle's Table 5 would be 342.)

The intercept coefficients C and w_0 may even be bypassed altogether by writing the two models for any two 'states' 1 and 2 of each system as

$$\begin{gathered} P_1V_1\simeq P_2V_2, \\ w_1(1-b_1)\simeq w_2(1-b_2). \end{gathered}$$

In any case, work subsequent to 1662 showed that C is related to the absolute temperature T, as in the gas law $PV\simeq RT$, where R is a constant. The value of R has in turn been found to be the volume occupied by one gram-molecule of the gas in

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

question, which is 22.415 l at $P = 760$ mm Hg and 0 °C (for a 'perfect' form of the gas). So C is no longer an arbitrary statistical number but is itself routinely predictable.

Similarly, w_0 can be thought of as w in the limit as b tends to 0 (i.e. w for a very small brand). Its value w_{0t} in any analysis period of relative length t is related to w in the chosen unit time period by the formula $w_{0t} - 1 \simeq t^{0.28}(w_0 - 1)$. (This is a good approximation to the Poisson–gamma model which in turn underlies the negative binomial distribution of Table 8—Ehrenberg (1959, 1988).) More generally, w_0 (and the DJ relationship itself) is subsumed by the Dirichlet choice theory of buyer behaviour (Goodhardt *et al.*, 1984; Ehrenberg, 1988), as is noted briefly in Section 6.5.

These remarks indicate the more advanced theoretical developments that can occur with the parameterization of such models, *once an observed relationship has become routinely predictable*. They also emphasize the similarity of the two case histories.

5.4.2. *Less well-developed cases*

We are not suggesting that most relationships which people study will work out as simply as this, but only that when routinely predictable or law-like relationships *do* emerge then they will often, we think, follow something like this route.

There are less well-developed cases of predictable models with arbitrary looking numerical slope coefficients. An example is the two models $\bar{h} \simeq 5.3 \log \bar{w} - 46$ for the mean heights $\bar{h}$ and weights $\bar{w}$ of Indo-European boys at 2–18 years of age, and $\bar{h} \simeq 51 \log \bar{w} - 39$ for all other boys worldwide and for all girls up to 13 years of age (e.g. Ehrenberg (1975) and Bound and Ehrenberg (1992)). Here the numerical slope coefficients 53 and 51 are each effectively constant and thus routinely predictable within a wide range of conditions (based on about 3000 different data sets for 200000 children of many ethnic origins in 42 countries).

There are also many relationships which have *not* become routinely predictable, or at least not yet quantitatively. An example of the latter is the DJ relationship for *attitudinal* data, e.g. people's beliefs about brands or politicians (see Table 6 and Section 6.5). Here full quantification has not (yet) been achieved, partly because the scales of measurement which are used differ in ways that are not (yet?) sufficiently well understood.

6. DISCUSSION

In this final section we try to place our main arguments concerning routine predictability in the context of

(a) ordinary science as we see it (Section 6.1),
(b) traditional statistical teaching (Section 6.2),
(c) Hume's paradox in the philosophy of science (Section 6.3),
(d) the traditional *ceteris paribus* assumption (Section 6.4),
(e) the supposed contrast between the social and physical sciences (Section 6.5), and
(f) some summary comments (Section 6.6).

6.1. *Routine Predictability in Science*

The aim in our case histories was not to predict any specific values of the variables, but the relationships $PV \simeq C$ and $w(1 - b) \simeq w_0$ between them, with near certainty. A

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

failure of such a 'near-certain' prediction in science does not mean that the basic result is then no longer routinely predictable. We can usually see afterwards what has gone wrong, in terms of the already known kinds of exceptions, and we can test for this by suitable repetition.

There could possibly have been a new reason or factor why the predicted result did not hold. If properly confirmed, this would be a discrepancy with all that has gone before—something *new*, a discovery worthy of a letter to *Nature*. But worthwhile scientific discoveries are seldom that easy. Once a result is empirically well grounded, serious discrepancies are rare. Most are more likely to be instances of Twyman's law: 'Any result that looks different or interesting is usually wrong'.

In practical life we often have to act not on firm predictions but on questionable forecasts, historical interpretations, or hunch. These are difficult to make correctly because sufficiently clear previous experience is lacking. When Allsop (1987) felt unable to predict the dollar exchange rate 3 years ahead, he rightly thought this to be difficult not because it was for the future but because 'the behaviour of the dollar had so far defied satisfactory explanation even with the luxury of hindsight'. Inability to forecast correctly is usually not, in our view, because the future is too uncertain (despite the mistaken popularity of this concept nowadays), but because, much more simply, we are as yet too ignorant about the past. (Why *was* the dollar exchange rate what it was in 1990?)

We do not believe that all phenomena can be made routinely predictable or 'scientific'. Many appear to be too nearly unique or irreproducible to study effectively in that way, or too complex. But without making a fetish out of 'chaos', being able to predict that an outcome is unpredictable (like heads or tails of a properly spun coin) seems also a valid, if mostly perhaps less useful, advance in our knowledge.

We also do not believe that establishing a routinely predictable result ('brute empiricism') is the ultimate goal of science. *That* consists of the much more complex processes of

(a) developing grounded theory and causal understanding of the observable phenomena (as is widely discussed elsewhere, e.g. Ehrenberg (1975), part VI, Ehrenberg (1982a), part V, Ziman (1991), etc.), as well as
(b) practical 'engineering-type' applications (e.g. Ehrenberg and Uncles (1992) for applications of DJ and related phenomena).

Routine predictability as discussed here is merely a necessary but not a sufficient condition of good science:

> 'Before a theory explaining a process can be tested, that process must be known' (Einstein, 1949).

6.2. *Routine Predictability in Statistics*

Most basic statistical teaching—at least as in our standard text-books—differs from normal science by simply not addressing the issue of routine predictability at all, even implicitly. Texts give us least squares regression to provide the 'best' prediction. However, they seldom if ever discuss what this means, let alone whether their best predictions are any good (Corlett, 1963). Even merely the notion that predictions should be tested is seldom pursued. Nor is there (not surprisingly therefore?) any backlog of routinely successful predictions.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

We have found no justifications for regression methods in the books other than unsupported one-sentence assertions like 'Least squares is best for prediction'. Most statisticians merely seem to hold these truths to be self-evident, that all men are endowed with certain inalienable rights, among them life, liberty, the pursuit of happiness, and least squares.

But in his major review more than 40 years ago, Lindley (1947) admitted that the method of least squares 'is not easily justified except in certain cases', though he did not say what these cases were or what was special about them. And although Kendall (1951) in another review paper at about that time claimed that 'there is something to be said for accepting the principle of least-squares in its own right' he also did not say what that something was. Anscombe (1967) in an apologia for regression was reduced to let-outs such as 'it seems desirable', 'a satisfactory solution', 'there is reason to think', 'it is natural to eliminate' and 'we need not be abashed'. He said that those who disapprove of least squares regression methods altogether have 'undoubtedly much good reason on their side', but he typically also failed to say who this was and what their reasons might be.

Gauss himself set the tone around 1800 (see Gauss (1823)) when he justified minimizing his *squared* deviations, rather than Laplace's mean deviation (although only a very minor technicality in our context here), on the grounds that 'Unless we are mistaken, this [his] choice is surely no less arbitrary than ours'.

None-the-less, the use of optimization techniques such as best fit least squares regression has become one of the two paradigms of modern (classical) statistics. It has created a self-feeding expectation of instant gratification: faced with *any* new SSOD, least squares always immediately gives us the supposedly best predictions.

However, bivariate regression generally gives *two* answers (e.g. Berkson (1950)). These cannot possibly both hold for any other, different, data, and hence in practice *neither* will (Ehrenberg, 1963). For example, for the store choice data of Table 7, the two least squares regressions are

$$w = 5.55b + 2.22,$$
$$b = 0.090w - 0.11,$$

(which in '$w = f(b)$' terms would read $w = 11.1b + 1.2$). Both equations give a good fit to the data that they are fitted to ('nowcasting'), but neither holds for any other data that we know of (e.g. Tables 1–3), i.e. neither equation is in any way predictive.

What is more, least squares regression does not even give a much better fit to the data in hand than would any reasonable alternative (Ehrenberg, 1982b). For example, the above regression of w on b with slope 5.55 for Table 7 fits those data within a (minimum) residual standard deviation of 0.49. But equations with slopes of either 2.75 or 8.25 (a ratio of 300%) have residual standard deviations only 10% higher, and still 0.5 to the nearest single decimal. Least squares regression seems a case of the highly questionably best being the enemy of the good.

There appears, however, to be a distaste in statistics for searching explicitly for a model which could hold across MSOD (e.g. Ehrenberg (1990)). Some of our colleagues (including one of the referees) have even told us 'That's not statistics, that would be *science*!'. Our statistical texts and journals are instead obsessed with purely deductive techniques of statistical inference (type (ii) predictions in Section 1), based on rather elementary mathematics. Yet as Etzioni (1984) has stressed in the context of economics: 'Above all, we need more induction'.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

We recognize that there is likely to be a gulf between what our text-books say and best (or even perhaps just normal) statistical practice. Applied statisticians probably behave more like ordinary scientists than our statistical literature leads us to expect. Some of what they do may even find its way into some teaching. But it is not well documented.

Some attempts to go beyond the traditional technique-orientated stance have been noted in Section 3.3. But they have not led to what we ourselves have been able to recognize as routinely predictable results or workable principles. If statisticians really want to change tack, they would need to change the design and organization of their studies (see for example Nelder's recent emphasis on *design*—News & Notes (1992)): any one study should seek to contain at least one well-differentiated replication.

6.3. *Routine Predictability and Philosophy of Science*

Questions of the routine predictability of a result have been widely discussed by philosophers of science, often touching on Hume's famous paradox of induction. This is to the effect that we cannot logically justify from the previous regular occurrences of the same event (however many) that it represents a universal law, i.e. that the event will invariably recur. Yet we seem to behave as if we could.

The paradox can, we suggest, be resolved by noting two adjustments (without necessarily aiming at full philosophical rigour here):

(a) the empirical evidence for induction consists not of a large number of 'identical' repetitions, but of a wide range of *varied* replications;
(b) the laws of science are conditional rather than universal, i.e. each holds under certain conditions and never under others.

The latter seems to follow if we accept that the laws of science are not strictly true anyway (e.g. that the gas laws $PV = RT$ hold only for perfect gases, that perfect gases are defined as substances for which the gas laws hold, and that in practice there are no such things—Uvarov *et al.* (1968)).

We believe that we can therefore rephrase the basic statement of empirically based induction as one of interpolation:

> given successful replications over a known range of conditions, one can predict that the event will recur yet again within that range of conditions.

Any serious exception—e.g. that the sun did *not* rise tomorrow—would imply a change in the stated conditions of observation, something new. It would rate as a scientific discovery (a mini-paradigm shift, or the end of the world as we know it), and not merely a failure to predict. It does not happen often or easily.

Some fuzziness remains, we think, over what precisely 'that range of conditions' means. For example, does the fact that $w(1-b) \simeq w_0$ holds for cooking oil and dentifrice in a Pacific rim country such as Japan also cover another like Australia? We believe that such a doubt exists only at the margin. It does not require any 'leap of faith' (e.g. Draper *et al.* (1993)) because it is readily testable. Given all the other accumulated evidence, a single successful test of $w(1-b) \simeq w_0$ for just one product in Australia would in principle be enough. Any subsequent failure of DJ there could then no longer be just because it is *Australia*.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

6.4. *The* Effects *of Other Things being Equal*

The assumption of *ceteris paribus* has been the crucial enabling concept in many theoretical arguments. But we would argue instead that it is only the *effects* (or *lack* of effects) of the other things that have to be known or assumed to be the same.

As already noted, there are always many 'other things' which are not (and should not) be the same in different empirical replications—it is a different time and/or place, a different grocery product or body of gas, we are older and wiser next time, the stars are in a different position (which matters to many people), there is still a recession again perhaps, and so on.

Instead of assuming that these undoubtedly different things are 'equal', the crux of our argument here is that variations in these factors are very real but do not matter because they have no noticeable *effects* on the law-like relationships in question, i.e. that Boyle's law or DJ still hold. The proof of the pudding is in the results.

6.5. *Is Social Science a Science?*

It is commonly said that phenomena involving human beings cannot be as predictable as phenomena in the physical sciences. To quote one of the referees:

> 'Is there not a serious difference between the theories available to support Boyle's Law and those for DJ? In the former, the physical theories give the law and many other results. There is no social science theory of such generality. Furthermore, no one would be surprised if DJ ceased to hold because of a change in social structure, whereas it is hard for anyone to conceive of a change in physical structure.'

But it seems to us that such critics of social science usually know little science (whether physical or social) and less of its history.

The modern theories to support Boyle's law did not exist *in his time*. His 'A defence of the doctrine touching the spring and weight of the air' (Boyle, 1662) was in direct opposition to the then current thesis of the *funiculus* (Linus, 1661). This was that nature provides hypothetical filaments or 'strings' of extremely rarefied matter to hold up the column of mercury in our now familiar mercury-in-glass barometer (the then famous Toricellean experiment). Boyle argued from his results against the *funiculus*

> 'wherefore since besides the several difficulties that incumber the hypothesis we oppose, and especially its being scarce, if at all, intelligible, we can add that it is unnecessary'.

The situation has not greatly changed with time (Ziman, 1991):

> 'The active research literature of physics is well supplied with fanciful conjectures [e.g. "superstrings"?] for which there is little evidence and which are later tacitly dropped overboard. . . . The physics of undergraduate text-books is 90% true, the contents of the primary research journals of physics is 90% false.'

In marketing, the social sciences and statistics the situation is generally still as yet somewhat worse (only marginally of course for our journals, though *very* much so for our text-books). But we were late starters and will no doubt be catching up.

The present-day explanatory theories of Boyle's law such as molecular movements in general, and the kinetic theory in particular, only came a hundred or so years later. And as for the alleged *permanence* of the physical sciences, black holes are nowadays claimed to represent a dramatic change in our perception of physical

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

structure, as did $E=mc^2$, Galileo's 'And yet she moves', superconductivity and almost innumerable other paradigmatic or near-paradigmatic shifts in physics in their time.

As for our particular social science illustration here, McPhee (1963) developed his theoretical explanation of DJ almost instantly once the empirical phenomenon was first noted by Jack Landis in the early 1960s. Expressed in terms of a current example for politicians (Ehrenberg, 1991; Mills and Ehrenberg, 1992), it was based on the assumption that whereas fewer people will have heard of one politician than of another, those who had heard of both would regard them as of equal merit. DJ then *has* to follow. Fewer of the few people who have heard of the less-well-known politician would rate him as their favourite compared with the larger proportion of the more numerous people who have heard of the better known politician and who rate him or her as *their* favourite.

The theoretical derivation of the fully quantified DJ relationship $w(1-b) \simeq w_0$ as in Section 5.4 came within less than 10 years, and the much wider ranging stochastic Dirichlet model as a mixture of Poisson, gamma, multinomial and beta distributions from which DJ follows as just one of many predictable results (the technical equivalent of the kinetic theory) came within less than 20 years (both due to our colleagues Professor Gerald Goodhardt and Dr Chris Chatfield—e.g. Goodhardt *et al.* (1984), Ehrenberg (1988, 1993), and earlier references there).

There could come a time when people will no longer have look-alike brands or politicians to choose between (although the demise of state communism may make this seem less likely rather than more). But when such commodity-like choice opportunities exist (as they always do when there is strong competition), then DJ will *have* to occur. In other words, whereas the *conditions* for DJ might just conceivably disappear down some social black hole, their theoretical DJ-type consequences will not.

It therefore seems to us that there is no fundamental difference between the social and physical sciences, either in terms of the occurrence of routinely predictable empirical patterns or of their underlying theories. This matters, both for the pursuit and practical applications of the sciences and for their funding.

6.6. *Summarizing Remarks and Conclusions*

We have argued in this paper that the routine predictability of an observed phenomenon depends on previous experience of its having consistently recurred under a wide variety of differing conditions. This requires observational and/or experimental studies which are designed to cover MSOD, observed under very different conditions. The appropriate analytical criterion is then whether the model holds for such MSOD, and *not* what model fits best for one particular SSOD (and no others).

The process is empirical, with only low level descriptive modelling. Deeper theory can none-the-less contribute, for example in

(a) indicating a suitable functional form,
(b) suggesting telling conditions of observations to vary,
(c) integrating disparate phenomena (e.g. the buying of brands and attitudes to politicians),
(d) pinning down the nature of exceptions,
(e) predictive extrapolations to quite different (but theoretically linked) circumstances and
(f) providing causal explanations and understanding more generally.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

We have sought to illustrate that different sciences need not differ in the degree and nature of their predictable phenomena (although the history of successful social science is much briefer). To us the odd man out is classical (i.e. 'modern') statistics. This seems to us to suffer from its obsession with distilling the most out of an isolated, and essentially also *small*, set of data (usually just to cope with the statistical uncertainty of a doubtful borderline result). It has failed to share the scientist's enthusiasm for describing and explaining widely occurring and ultimately routinely predictable phenomena.

Classical statisticians have almost totally ignored that we can fit a line to more than one set of data, and that we should then extend the predictive validity of any such relationship to yet further, and different, data sets. The criterion of 'How far does it generalize?' should, we think, replace 'Is it significant?' as the analyst's real touchstone (rather than the often suggested but purely applied and situation-specific notion of 'practical importance', which has nothing to do with objective knowledge). And as for Gauss's least squares, besides its being (to us) 'scarce', if at all, intelligible (i.e. why the slope of a predictable relationship should in any way be determined by the scatter of the more or less random errors about it, rather than the other way round), we can add that (with MSOD) it is unnecessary.

None of this is to deny the value of a statistical approach, though mainly in the broader and perhaps more demanding sense of analysing MSOD, implying coping with large cumulative samples, and with little call for techniques of probabilistic inference and optimization. The underlying theories in both our case histories are, however, unashamedly stochastic, so that there is still plenty of room for probabilistic ideas, but at broadly descriptive and explanatory levels rather than at a narrowly *inferential* level (and probably dealing with phenomena which themselves are being assumed to be quasi-random, rather than expecting deep pay-offs from data where only the *residual errors* are made or assumed to be random).

There is also scope for mathematics. But it would be more of the kind which is able to show why the DJ phenomenon has to occur (e.g. as a selection effect like Simpson's paradox) or why the assumptions underlying the broader Dirichlet model have to be what they are (Goodhardt and Chatfield (1973) and Goodhardt *et al.* (1984), section 2.2), rather than merely producing yet another maximum likelihood estimator.

Many statisticians on both sides of the Atlantic see our subject as undervalued by others and even suggest that science is therefore much less efficient and/or effective than it might be (e.g. Zellner (1992) and News & Notes (1992) currently.)! We ourselves believe instead that, if statistics were to acquire a radical concern for producing routinely predictable results, it would come more into line with the kinds of work that scientists, social or physical engineers, and also managers, actually do.

ACKNOWLEDGEMENTS

This paper is part of a programme of work that has been supported by Colgate Palmolive, New York, and leading companies in the UK. We are indebted to comments from Neil Barnard, Patrick Barwise, Chris Chatfield, John Nelder, David Targett, John Tukey and particularly also Helen Bloom, as well as from many other colleagues.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

REFERENCES

Aitchison, J. A. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis*, p. 1. Cambridge: Cambridge University Press.

Allsopp, C. J. (1987) The rise and fall of the dollar: a comment. *Econ. J.*, **97**, 44–48.

Anscombe, F. J. (1967) Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *J. R. Statist. Soc.* B, **29**, 1–52.

Aske Research (1969) *Loyalty Reports: Ready-to-eat Cereals*. London: Aske Research.

Barnard, N. R., Ehrenberg, A. S. C., Geroski, P. A. and Kau, A. K. (1993) Brand loyalty and market structure. *CMaC Working Paper*. London Business School, London.

Bartlett, M. S. (1949) Fitting a straight line when both variables are subject to error. *Biometrics*, **5**, 207–212.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. New York: Springer.

Berkson, J. (1950) Are there two regressions? *J. Am. Statist. Ass.*, **45**, 164–180.

Bound, J. A. and Ehrenberg, A. S. C. (1992) Model extension and model tuning. Submitted to *Appl. Statist.*

Boyle, R. (1662) A defence of the doctrine touching the spring and weight of the air. In *New Experiments*, 2nd edn. London.

Christie, A. A. (1990) Aggregation of test statistics. *J. Accntng Econ.*, **12**, 15–36.

Cobb, G. (1991) Teaching statistics: more data, less lecturing. *Amstat News*, **182**, 1, 4.

Cohen, I. B. (1964) Newton, Hooke, and 'Boyle's Law'. *Nature*, **204**, 618–621.

Copas, J. B. (1983) Regression, prediction and shrinkage. *J. R. Statist. Soc.* B, **45**, 311–335.

Corlett, T. (1963) Ballade of multiple regression. *Appl. Statist.*, **12**, 145.

Cuthbertson, K., Hall, S. G. and Taylor, M. R. (1992) *Applied Econometric Techniques*. Ann Arbor: University of Michigan Press.

Deming, W. E. (1992) Confidence intervals. *News & Notes*, **19**, no. 1, 6.

Draper, D., Hodges, J. S., Mallows, C. L. and Pregibon, D. (1993) Exchangeability and data analysis (with discussion). *J. R. Statist. Soc.* A, **156**, 9–37.

Ehrenberg, A. S. C. (1959) The pattern of consumer purchases. *Appl. Statist.*, **8**, 26–41.

——(1963) Bivariate regression analysis is useless. *Appl. Statist.*, **12**, 161–179.

——(1972) *Repeat-buying*. London: Griffin.

——(1975) *Data Reduction*. New York: Wiley.

——(1982a) *A Primer in Data Reduction*, ch. 13. New York: Wiley.

——(1982b) How good is best? *J. R. Statist. Soc.* A, **145**, 364–366.

——(1988) *Repeat-buying*, 2nd edn. New York: Oxford University Press.

——(1990) A hope for the future of statistics: MSoD. *Am. Statistn*, **44**, 195–196.

——(1991) Politicians' double jeopardy: a pattern and exceptions. *J. Mkt Res. Soc.*, **33**, 347–353.

——(1993) Double Jeopardy: a review of the market. *Nature*, to be published.

Ehrenberg, A. S. C., Goodhardt, G. J. and Barwise, T. P. (1990) Double jeopardy revisited. *J. Mktng*, **54**, 88–91.

Ehrenberg, A. S. C. and Uncles, M. D. (1992) Management applications of the Dirichlet choice model. Submitted to *J. Mktng Res.*

Einstein, A. (1949) *Out of My Later Years*. New York: Philosophical Library.

Eisenhart, C. (1946) The assumptions underlying the analysis of variance. *Biometrics*, **3**, 1–21.

Ellis, K. (1989) Private label brands. *PhD Thesis*. London University, London.

Ellis, K., Hammond, K. A., Barnard, N. R. and Ehrenberg, A. S. C. (1993) Loyalty for private labels. *CMaC Working Paper*. London Business School, London.

Etzioni, A. (1984) *The Moral Dimension: Towards a New Economics*, p. 19. New York: Free Press.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

——(1926) The arrangement of field experiments. *J. Min. Agric.*, **33**, 503–513.

——(1935) The logic of inductive inference (with discussion). *J. R. Statist. Soc.*, **98**, 39–82.

Fuller, W. A. (1987) *Measurement Error Models*. New York: Wiley.

Gauss, C. F. (1823) *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. Gottingen: Dieterich.

Geary, R. C. (1943) Accuracy of Boyle's original observations on the pressure and volume of a gas. *Nature*, **151**, 476.

Goodhardt, G. J. and Chatfield, C. (1973) Gamma distribution in consumer purchasing. *Nature*, **244**, 316.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Goodhardt, G. J., Ehrenberg, A. S. C. and Chatfield, C. (1984) The Dirichlet: a comprehensive model of buying behaviour (with discussion). *J. R. Statist. Soc.* A, **147**, 621–655.
Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman and Hall.
Hedges, L. V. and Olkin, I. (1985) *Statistical Analysis in Meta-analysis*. Orlando: Academic Press.
Judge, G. G., Griffiths, W. E., Hill, R. C., Luthepohl, H. and Lee, T. C. (1985) *The Theory and Practice of Econometrics*, 2nd edn. New York: Wiley.
Kau, A. K. (1981) Patterns of store choice. *PhD Thesis*. University of London, London.
Kendall, M. G. (1951) Regression, structure, and functional relationship, I. *Biometrika*, **38**, 11–15.
——(1952) Regression, structure, and functional relationship, II. *Biometrika*, **39**, 96–108.
——(1961) Natural law in the social sciences. *J. R. Statist. Soc.* A, **124**, 1–16.
Kendall, M. G. and Stuart, A. (1979) *The Advanced Theory of Statistics*, 4th edn, vol. 2. London: Griffin.
Kennedy, P. (1991) *A Guide to Econometrics*, 3rd edn. Oxford: Blackwell.
Leftwich, R. L. (1990) Aggregation of test statistics. *J. Accntng Econ.*, **12**, 37–44.
Lindley, D. V. (1947) Regression lines and the linear functional relationship. *J. R. Statist. Soc.* B, **9**, 218–244.
Lindsay, R. M. and Ehrenberg, A. S. C. (1992) The design of replicated studies. *Am. Statistn*, to be published.
Linus, F. (1661) *De Corporum Inseperabilate*.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
McPhee, W. N. (1963) *Formal Theories of Mass Behaviour*. New York: Free Press.
Miller, A. J. (1990) *Subject Selection in Regression*. London: Chapman and Hall.
Mills, P. and Ehrenberg, A. S. C. (1992) Politicians' double jeopardy after the election. *CMaC Working Paper*. London Business School, London.
Nelder, J. A. (1986) Statistics, science and technology. *J. R. Statist. Soc.* A, **149**, 109–121.
——(1989) Private communication.
News & Notes (1989) The new curricula. *News & Notes*, **15**, no. 9, 1.
——(1992) Inefficient science. *News & Notes*, **18**, no. 8, 1–2.
Robinson, G. K. (1991) The BLUP is a good thing: the estimation of random effects. *Statist. Sci.*, **6**, 15–51.
Smith, T. M. F. (1991) Post-stratification. *Statistician*, **40**, 315–323.
Tukey, J. W. (1989) The philosophy of multiple comparisons. 1989 Miller Memorial Lecture, Stanford University.
Uncles, M. D. and Ehrenberg, A. S. C. (1990) The buying of packaged goods at US retail chains. *J. Retlng*, **6**, 278–296.
Uncles, M. D. and Hammond, K. A. (1992) Store visits: a model-building approach. *CMaC Working Paper*. London Business School, London.
Uvarov, E. B., Chapman, D. R. and Isaacs, A. (1968) *A Dictionary of Science*. London: Penguin.
Wachter, K. W. and Straf, M. L. (1990) *The Future of Meta-analysis*. New York: Sage.
Wald, A. (1940) Fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.*, **11**, 284–300.
Wrigley, N. and Dunn, R. (1984) Stochastic panel-data models of urban shopping behavior. *Environ. Planng* A, **16**, 629–650, 759–778.
Zellner, A. (1992) Statistics, science and public policy. *J. Am. Statist. Ass.*, **87**, 1–6.
Zellner, A. and Hong, C. (1989) Forecasting international growth rates using Bayesian shrinkage and other procedures. *J. Econometr.*, **40**, 183–202.
Ziman, J. (1991) *Reliable Knowledge*. Cambridge: Cambridge University Press.

DISCUSSION OF THE PAPER BY EHRENBERG AND BOUND

R. B. Davies (Lancaster University): Professor Ehrenberg and Mr Bound have come before the Society as heretics, rejecting statistical orthodoxy as inimical to science. I look forward to some heated discussion but progress requires that orthodoxy be challenged and I am sure that many of us welcome their paper in this spirit.

For my part, I particularly welcome the emphasis on replication. I agree that there is a tendency in the social sciences to claim definitive results from analyses of single sets of data. Sometimes this is a

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

deliberate exaggeration of projects, thought to be necessary to obtain funding or publication. However, it must also be recognized that a synthesis of the results of studies based on observational data is highly problematic because of the different effects compounded in different data sets and different analyses.

I also welcome any encouragement to take model formulation more seriously: this is a weak area in the social sciences. Even in econometrics, with a well-developed tradition of formally deriving models from theoretical postulates, derivations can be vulnerable to accusations of tautology with error assumptions often made to ensure that the reduced form model has a conventional specification.

However, I take exception to two fundamental features of this paper. First, I believe that the authors are far more sanguine about the merits of pursuing simple quantitative axioms than experience in social science to date would warrant. Second, 'prediction' and 'predictability', as defined in this paper, are only one of a plurality of objectives justifiably addressed by empirical work in the social sciences. I am readily persuaded that orthodox statistics has a limited contribution to make in the context developed by the authors, but I see no logic in generalizing from this context to draw conclusions about empirical endeavour in social science as a whole.

There have been many attempts to develop simple axioms in the social sciences. A few appear to have been successful in a manner closely analogous to Boyle's law, with the initially empirically estimated parameters eventually having some sound theoretical rationale. I am grateful to Professor Graham Hitch for drawing my attention to one example: the articulating loop hypothesis. This is the relationship between speech rates (words per second) and memory span (measured by the mean number of words recalled). The relationship is linear for children 4 years of age up to adults and for different vocabularies (Hulme *et al.*, 1984). It has been tested for English, Arabic, Chinese and Welsh. The intercept varies with different languages, but the slope remains the same at 1.5 s and is now thought to measure a fundamental feature of human memory. Moreover, as emphasized to be important by Ehrenberg and Bound, deviations from this simple relationship have been probed by varying the experimental conditions and theoretical explanations have been developed for the results. All these achievements have required little statistical input.

However, not only are such examples of 'predictability' rare in the social sciences, but success in developing a simple model may be illusory. Consider the McGinnis (1968) axiom of cumulative inertia. This claims great generality for a very simple relationship:

> 'The probability of remaining in any state of nature increases as a strict monotone function of duration of prior residence in that state'

—be it marriage, job or almost anything except life itself (for which it does not hold). This axiom and other 'social physics' relationships suggested in the 1960s rate highly on parsimony and generality but I have yet to find a social scientist who is impressed. I believe that they are correct not to be impressed because this type of aggregate relationship achieves its generality by averaging over the myriad of systematic features of behaviour which are of prime interest to social science researchers.

This could be characterized almost as a dual of Simpson's paradox: that substantively interesting systematic relationships may be eliminated by aggregating to a level at which a stochastic banality is achieved. Where in the McGinnis axiom is there any appreciation of the myriad of factors involved in retaining a job, moving through the life-cycle, obtaining training, the constraints imposed on and opportunities that befall an individual, and so on, and how these affect movement from job to job, or, in residential tenure, how the different factors which can either limit the ability or increase the opportunities to move evolve over time? The declining hazard may not even be meaningful; it may be a spurious relationship due to aggregating over heterogeneous samples. There is extensive empirical support for the McGinnis axiom but it contributes little to substantive social science. I wonder whether the double-jeopardy example has the same banality.

This paper does not reflect the excitement and challenge that I enjoy in working with social scientists to disentangle the relationships within complex data. Nor does it reflect the even larger challenge of developing models which may be used to predict the effect of different settings of policy instruments. To suggest that attention should be confined to models with just one or two parameters ignores much of the reality of social science investigation and trivializes the problems to be confronted in developing methods for evaluating a range of policy instruments. But I acknowledge that plurality of approaches is one of the hallmarks of social science.

I would propose the vote of thanks to the speakers.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

J. B. Copas (University of Warwick, Coventry): A strength of our Society and these meetings is that we have papers representing a great variety of views of our subject, and tonight we are invited to rethink our approach to prediction. The authors make the case that we give insufficient attention to developing models which generalize, and too much attention to fitting models to small samples: very sensible, but we need to know more, and so we look to the paper for guidance. This is where our problems begin.

Firstly, we have to be clear about the way that we use technical terms. Those of us involved in Science and Engineering Research Council affairs will be familiar with the new term 'complex stochastic system'. This term is a technical device for extracting research funding for statistics. When asked what a complex stochastic system is, however, the chairman of the relevant committee, a leading Fellow of our Society, seemed somewhat taken aback by the question, but thought that he would recognize one if he saw one. I am not sure that I recognize some of the terms as used by the authors in this paper. Take prediction for example. By prediction most of us have in mind making a statement about an as yet unobserved event. Modelling the uncertainty in this event is the special feature of *statistical* prediction. But the paper speaks of 'predicting the relationship between variables'—more a matter of data summary or description. To criticize the many important developments that have been published on the grounds that they tackle a different problem, I find unconvincing, and even more so when the contention is that that problem is trivial or irrelevant.

Frequent reference is made to the distinction between single sets of data and many sets of data—but what is really meant here? Fig. 2 shows observations on two variables w and b, with the data falling into four groups. Is this a single set of data, four sets of data or, as each point is an average, perhaps 20 sets of data? Whatever it is, it is clear that each group is nicely linear, giving a two-parameter family of straight lines. By making the lines go through a common point, we obtain a one-parameter family of straight lines. The graph shows the data in the first four tables of the paper, and the authors' model is a one-parameter family of hyperbolae. The data themselves give no evidence whatsoever for curvature, the choice of the double-jeopardy (DJ) model apparently resting on the rather dubious assumption of independence, that whether I buy a packet of Rice Krispies is independent of whether I have just bought a packet of Corn Flakes. Does this imply a lack of memory property of the distribution of breakfast cereal? I prefer a constant amount of breakfast cereal each day, so I shall stick to my straight line

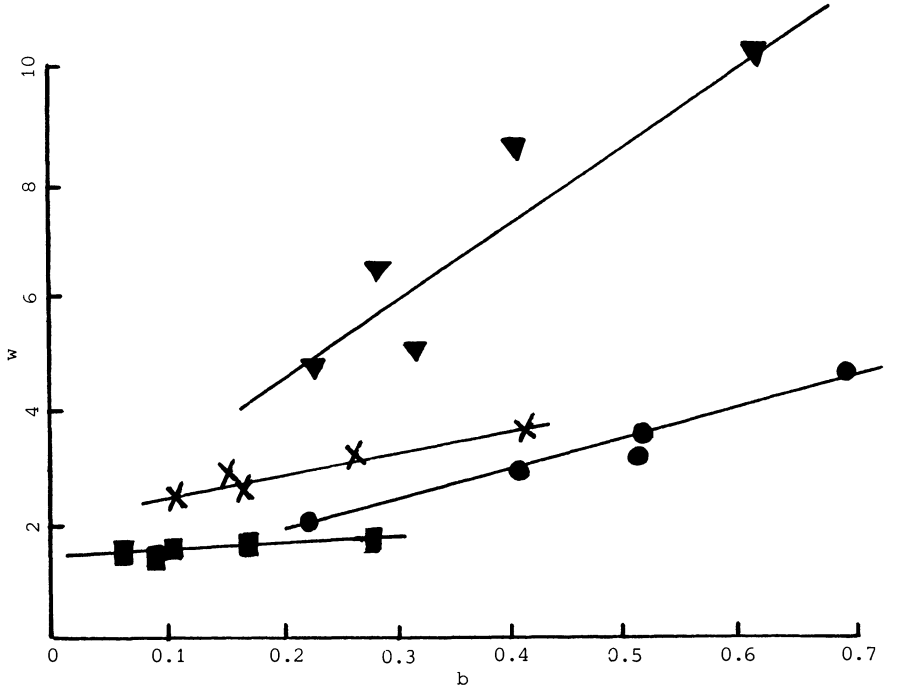

Fig. 2. Data from Tables 1–4 of the paper

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

model—it is simpler and fits better. Incidentally, the graph seems a much more effective way of displaying the data than the separate tables in the paper.

Surely the important thing about these models is that they have an adjustable parameter that can be separately fitted to each data set. To say that the DJ model is better than the line considered in Section 2.1 is absurd—we are comparing a single line with a family of curves. Boyle does much better, as his adjustable parameter C can itself be modelled in terms of the gas constant, earning the status of a scientific law with flying colours. I would be surprised if there was any comparable way of modelling w_0.

A major worry with the paper is the almost total lack of attention to sampling errors or uncertainty. Our subject has grown into what it is today because of the many imaginative ways of modelling uncertainty that have been developed, whether expressed in terms of natural variability, measurement error, randomization or sampling from real or imagined populations. These concepts should not be lightly dismissed. Measurement errors *are* crucially important in prediction. Table 5 suggests that Boyle decided to adjust the volume to a convenient set of values, and then to read the pressure from his apparatus. Inevitably there will have been errors. Suppose that for nominal volume v the actual volume is $v+\epsilon$, where ϵ is a random error with mean 0 and standard deviation σ. Then the expected pressure is

$$E\left(\frac{c}{v+\epsilon}\right) = \frac{c}{v}\left(1+\frac{\sigma^2}{v^2}\right) + o(\sigma^2).$$

Fig. 3 shows the data of Table 5, a beautiful fit to a rectangular hyperbola. This formula with $\sigma = 2$ (an unrealistically large value for Boyle, but perhaps not too unrealistic for other applications) gives

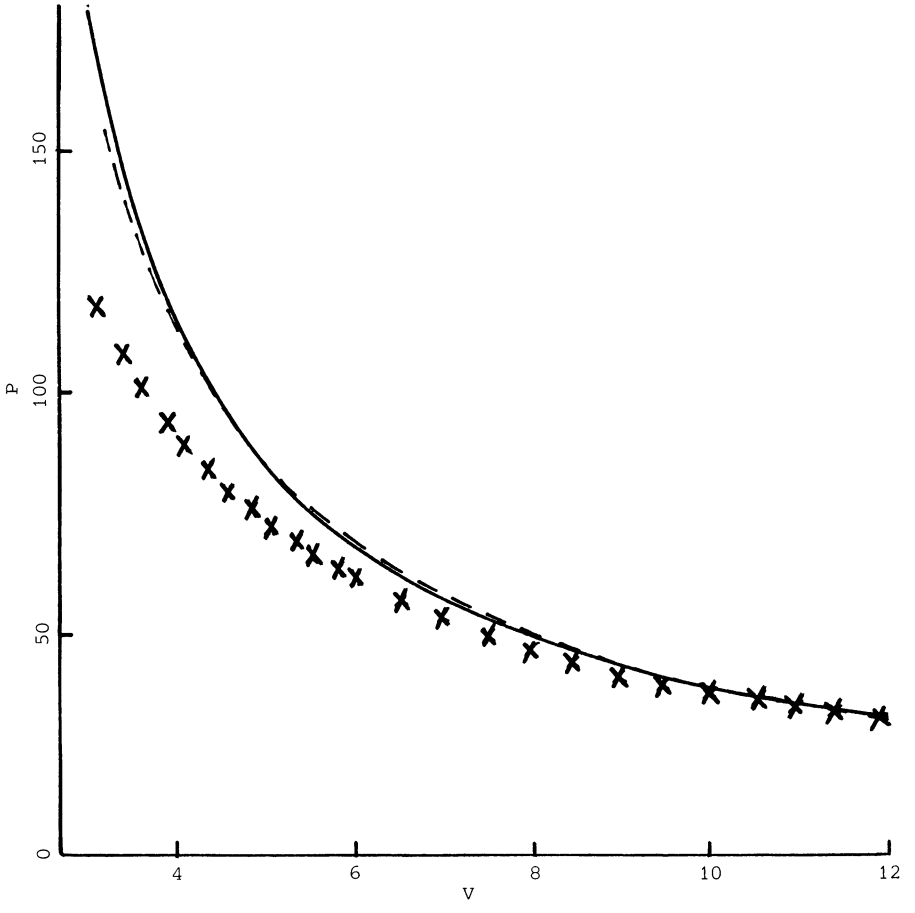

Fig. 3. Data from Table 5 of the paper: **X**, Boyle's data; ———, expected P; – – –, power law approximation

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

a quite different curve, close to $596v^{-1.2}$ over the range considered. So if the errors are ignored by taking averages, which seems to be what the authors might suggest, the wrong coefficient in the power law would result. The correct power of -1 could be recovered by modelling the data correctly as a non-linear function relationship.

The authors say that there is not a single quotable success story of statistical prediction. Many of us will be anxious to prove them wrong, so perhaps we can institute a prize for the best example. My entry is the reconviction prediction score used by the Parole Board in considering prisoners for parole. This statistical predictor, estimating the chance of reconviction within 2 years of leaving prison, was fitted to data in 1965 and has more recently been reassessed in Ward (1987). Although there has been some drift in reconviction rates, the predictor remains remarkably effective, despite a giant upheaval of the system in the intervening years with the introduction of parole. But perhaps this is not prediction in the sense used in the paper. My entry for the authors' alternative prize is the humble straight line which, with two adjustable parameters, gives a useful description of say 90% of data sets. Allowing a little licence with transformations this increases to perhaps 99%: a success story indeed.

The authors cite the rising of the sun as a routinely predictable event, and, after this evening, the sun will, God willing, rise again tomorrow. Similarly I believe that techniques of statistical prediction are routinely useful, and, after tonight's paper, will continue to be so tomorrow. On this note of confidence I have pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

C. Chatfield (University of Bath): I welcome this paper for focusing attention on many sets of data (MSOD) rather than on the single sets of data (SSOD) discussed by most statistics publications. Traditional statistical inference considers estimating and testing the parameters of an *assumed* model for a *single* set of data. Yet in practice model building is an iterative interactive process involving model formulation, parameter estimation and model checking. These three steps should be based on different data sets but in practice are often based on the same set. Then much of the theory developed for the classical problem is invalidated, although practitioners typically are unaware of this.

Where I part company with the authors is when they say, in Section 1, that there is not 'a single quotable success story' for a statistical prediction from an SSOD. In the authors' terminology this may be true as one cannot by definition generalize from an SSOD. However, 'prediction' is more often used to mean some sort of extrapolation. In time series one often only has a single series of a given type, but forecasts (extrapolative predictions in the terminology of the authors) still have to be made. The authors may regard this as undesirable or impossible, but it has to be done. Some such forecasts have proved reasonably 'successful' though I would prefer not to classify predictions as 'right' or 'wrong' but as 'more accurate' or 'less accurate'. This applies to both MSOD and SSOD predictions, and the good forecaster (or scientist) will give more accurate forecasts (or predictions) than a poor one.

Thus the statistician must be able to cope with both MSOD and SSOD. In the latter case, we should try to incorporate prior information and use procedures which allow for model uncertainty. If we already have a well-established model then it is easy to check new data against it and to tune or extend the model as appropriate. Rather little statistics is needed here. But setting up the model in the first place is what is difficult.

This paper gives little general guidance on the difficult task of model building. The remarks on between-group analysis assume that one has several samples of bivariate data and would not have helped to derive Boyle's law or the double-jeopardy law. Perhaps we should leave contextual modelling to appropriate scientific experts. Where the non-statistician *does* need help is with collecting data, with analysing an SSOD and with assessing uncertainty. Such work may not immediately give the lasting generalizable results sought by the authors but does play a vital role in the early stages of the scientific process and in many practical problems. Thus, while welcoming some aspects of this paper, I also believe that statisticians have a valuable role to play in the analysis of SSOD.

R. Fildes (Lancaster University): Ehrenberg and Bound's critique consists of two components:

(a) an approach to be used in developing 'scientific' laws;
(b) they argue that replication across a wide range of circumstances is critical in establishing the proposed laws and associated boundary conditions as true.

The paper's novelty lies in the completeness of the critique that the authors level at conventional statistical methods. The heart of the authors' criticisms is that

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

(i) 'inductive interpolative prediction' and its concern with multiple sets of data (MSOD) is the major interest of those wishing to use statistical methods in establishing predictive laws,
(ii) classical statistical methods are ill equipped to deal with this problem and
(iii) statisticians whose primary responsibility is to develop and test methods of analysis that illuminate important problems concerned with data analysis have spent little time in analysing this class.

Although mentioning various established methods that are applicable to the MSOD problem they unfortunately neglect their importance.

'Inductive interpolative prediction' despite the authors' implicit claim is closely aligned with the problem of forecasting (inadequately defined here by the authors) and both in turn are a subset of the problem of prediction outside 'conditions already covered'. They argue that the research need is to establish common models from which routinely predictable results across space and time can be derived but they fail to address questions of model adequacy and when a model could be said to fail. Nor do they attempt to justify their argument. In my field of business forecasting with a history of grandiose claims that have not lived up to expectations the more modest aims of statistical research are to establish methods which identify common data structures temporarily stable across time that may be cross-sectionally specific and to identify such approximations *ex ante*. Here forecasting accuracy is more important than the simplicity of a cross-sectionally applicable model and the authors wrongly deny the importance of unexpected future changes where past conditions are no guide.

However, the central charge that Ehrenberg and Bound make, that statisticians spend too much time on irrelevant problems, is true if we examine time series forecasting (Fildes and Makridakis, 1993). A content analysis of papers published in the past 20 years in the *Journal of the American Statistical Association* supports the authors' claims that applying the same model to populations which may be distinct has not been of concern to statistical researchers. In highlighting the critical issue of the relationship between data analysis, model building, model testing and the scientific method this paper is to be welcomed. Its weakness lies in a casual disregard for rigour in its definitions and the unstated and unexamined assumption that important statistical problems all have the same requirements.

S. Rosenbaum (Radlett): The two models in Section 5.4.2 appear to be a refinement of the single model said in Bound and Ehrenberg (1989) to hold for both sexes. Obviously the results are satisfyingly similar, and the conclusion must be that growth during childhood does follow a law-like relationship; T. J. Cole, in his comments on their paper, pointed out though that Quetelet's index w/h^2 is equally universal in its application: in logarithmic terms $\ln w = 2 \ln h + \text{constant}$.

By definition an adult has ceased to grow in stature but rotundity will most likely increase. In these circumstances the Quetelet index has proved its worth in the assessment of individuals, who are increasingly overweight in the UK these days, and it could be useful for monitoring groups of the population such as the armed services whose physique and fitness are important in their work. It can also provide a historical perspective.

Height, principally, has been held to represent the nutritional status of a country's population and has been correlated over time with its standard of living (Floud *et al.*, 1990) and other economic and health factors (Jordan, 1993); the latter work makes use also of the Quetelet index but admits that weight is comparatively scarce in Victorian data sets. Rosenbaum and Crowdy (1992) published a study of the average heights and weights of British Army recruits in every year from 1860 to 1974 that the data were available. Plotting height against weight results in a manifest curve which is still there to some extent when logarithms are taken of w and when $\ln w$ is plotted against $\ln h$. In reporting the following figures I take the view that height is the primary variable and weight is dependent on it and on other factors such as diet which are controllable. This justifies the calculation of regression slopes which I now give.

Over the whole period the slopes of $\ln w$ on $\ln h$ are 2.4 and 2.8 at ages 18 and 20–24 years respectively. But for average heights greater than $67\frac{1}{2}$ in (171.4 cm), coinciding with the period after the Second World War, the slopes are 4.4 and 5.9 at these ages. Clearly there has been a secular change of considerable proportions. The steepness of the trend over the latter period does not, however, invalidate the use of the Quetelet index for a particular subset of recruits. In 1951 national servicemen aged 18 had a slope, $\ln w$ on $\ln h$, of 1.8, and in 1980 a national sample of males aged 16–19 and 20–29 years (Knight,

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

1984) yielded values approximately equal to 2.3 and 2.1. The phenomenon appears to be a case of between (years) and within groups.

Peter Urbach (London School of Economics and Political Science): I share Ehrenberg's and Bound's criticism of standard regression methods as quite an unjustifiable system of inference and applaud their attempt to reintegrate statistical regression analysis with the inductive methods of the rest of the natural sciences, and especially their emphasis on the need for variety in evidential data. But their own approach is as much in need of justification as the approach that they criticize.

They rightly see Hume's problem of induction as central. Hume noted that, for example, the observation of 17 white swans does not entail any law about the colour of swans, nor any prediction about the colour of the 18th swan to be observed. Yet scientists regularly make such predictions: how?; and by what right? This is the problem of induction.

This is the authors' solution to the problem: 'Given successful replications over a known range of conditions, one can predict that the event will recur within the same range of conditions'.

The first problem is what does it mean to say that 'one can predict' some event *a*? Does it mean that 'it is physically possible to say in advance that event *a* will occur'? Clearly not—all events are predictable in that sense. Does it mean that 'event *a* will *definitely* occur'? Again, clearly not—this is what Hume showed. Presumably, when you predict some event, you are associating with it some (high) level of probability or certainty; if so, the inductive methodology that naturally suggests itself is the Bayesian approach, which I believe would shed light on the problems of regression and prediction. At any rate, in a paper on predictability and prediction these terms surely need to be defined.

The second problem is: when are two ranges of conditions 'the same'? Clearly they are never *exactly* the same, so I presume that the authors envisage predictability only when the conditions are *relevantly* similar. But they do not explain how relevance is to be determined. They give an example though. They ask whether, with respect to double jeopardy, Australian conditions are (relevantly) similar to Japanese conditions. This, they say, can be determined by checking to see whether the phenomenon actually occurs in Australia. But then the authors' position amounts to the claim that an event is predictable just in case it occurs, which is surely not what they intended, nor what could possibly suit the problems at hand.

A. J. Mayne (Milton Keynes): I found the paper and the discussion absolutely fascinating and tremendously stimulating. I do not yet know how far I agree with what has been said, but I would like to bring in some additional aspects.

The 'butterfly effect' in chaos theory emphasizes that processes like some forms of weather, which are very sensitive to initial conditions, become completely unpredictable after a fairly short time. However, in statistical mechanics, very complicated molecular systems, with many random variations, show very great regularity with very predictable behaviour on the large scale. Somewhere in between is 'antichaos', described for example by Kauffman (1991) and the Channel 4 Television Equinox programme shown on October 18th, 1992. Here, complicated systems with partly quasirandom behaviour nevertheless display some striking regularities.

It would also be interesting to explore the relationships between predictability, prediction and such recently considered causal modes as 'fuzziness' and 'fuzzy uncertainty', and 'coincidence' and 'acausal synchronicity' (Jung, 1972). In his Letter from America on November 15th, 1992, Alistair Cooke mentioned that no statistician whom he had met could explain the remarkable coincidences linking the late US Presidents Abraham Lincoln and John F. Kennedy.

A very striking feature of the universe that we know is that 'phenomena space', of which space–time is only a part, has large regions that are very regular and thus very predictable, interspersed at irregular intervals by regions of irregular shape that are very irregular and thus highly unpredictable. Both scientific method and statistical inference will have to be overhauled completely to cope with this! Yet we have to live with it in everyday life; we need to find out how to forecast those practical aspects of life that can be predicted reliably, but we *know* that many aspects of our life will *remain* totally unpredictable.

It seems to me that the British research councils, which cover mostly 'well-behaved predictable' phenomena, need to be supplemented by a 'Special Research Council' to sponsor explorations of the 'irregular unpredictable' realm!

The following contributions were received in writing after the meeting.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Richard Colombo (New York University): It is usually more difficult in the social and perhaps biological sciences to exert control over variables of interest than it is in the physical sciences. Consequently much of empirical social science is largely concerned with identifying variables and contingent factors that are relevant to a phenomenon. When regression analysis is used it is not usually to find a routinely predictable relationship but more modestly to find a routinely predictable set of relevant variables. On this admittedly weak test there are perhaps many regression success stories.

That more ambitious goals are not often attempted may be because statisticians have failed to develop good practical methods for analysing many sets of data (MSOD) (although the skill, effort, luck and insight involved in collecting and analysing MSOD to establish routine predictability should not be underestimated). In marketing, Professor Ehrenberg and his colleagues have successfully established routinely predictable relationships—so it can be done but there are few, if any, other examples. Much of the problem lies in the concern that statistics has with single sets of data, as the authors point out. This legitimizes studies based on one data set to the detriment of well-tried methods of science involving MSOD. For an example of the unsuitability of 'conventional' statistical approaches for tackling MSOD see Colombo *et al.* (1993).

The authors have showed how MSOD can be analysed when a small number of variables is involved. Where there are more than two or three variables, as is typical in social science, the search for a good model to fit MSOD is likely to be much more difficult. What are we to do in these circumstances?: settle for qualitative statements?; report regression coefficients for the data set but with a warning about the likely absence of predictability? Since much of social science leans heavily on statistical methods, statisticians have an opportunity to influence, perhaps radically, the way that social science research is conducted by developing techniques for MSOD. The plea for statisticians to be more 'scientific' and 'relevant' has often been made. With their focus on routine predictability and MSOD the authors eloquently repeat this plea but also offer a specific prescription about how it might be achieved. This paper should be widely read and pondered.

David Cox (Nuffield College, Oxford): The importance of studying relationships under a known range of different conditions is stressed in the traditional work on the design of experiments. See, for example, Yates and Cochran (1938) on variety trials. The insertion of factors into a design deliberately to achieve a good range of validity is standard practice.

There is much in the present paper with which to agree, but the following remarks concentrate on an issue on which tentatively I disagree strongly with the authors. In some sense science is about 'understanding what is happening'. Whatever the history, Boyle's law as an empirical prediction equation seems of fairly limited interest; its importance lies more as a link in a coherent set of ideas in classical physics involving thermodynamics, the kinetic theory of gases, the virial expansion and so on. Section 6 of the paper addresses these issues. Although a healthily empirical attitude to some of the more imaginative aspects of, for example, modern physics is no doubt desirable, it seems to me that the authors' account massively undervalues theoretical discussion. The authors see as the main challenge the development of more and better empirical prediction equations. An alternative would see the need for analyses that are more strongly based in the subject-matter knowledge of the field, i.e. involving substantive rather than empirical models. Although the balance must depend on the context, I see the second as the more pressing issue.

Finally there is not space to comment on the remarkably pessimistic statement in the final sentence of the second paragraph of Section 1.

D. R. Draper (University of California, Los Angeles) **and C. L. Mallows** (AT&T Bell Laboratories, Murray Hill): We like the paper's emphasis on prediction. However, the authors' main points provide new support for the old maxim that statisticians working on quite different types of applied problem may come to quite different conclusions about what is important and what has been neglected in the subject. If Professor Ehrenberg and Mr Bound had developed expertise in the field of medical research, say, where prediction of individual patient level outcomes is at the heart of comparison of alternative treatment protocols, rather than consumer research, where prediction of aggregate buying habits seems important, a paper like this might have had two different emphases: the authors would perhaps not have come down so hard on least squares and allied methods, which are about the best that we can often do when working with individual level data, and they would probably have noted that it is much easier to find law-like relationships at the aggregate level. We suspect that the statistical tent is sufficiently large to find room for workers searching for predictability at a variety of levels of aggregation.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Professor Ehrenberg and Mr Bound continue to castigate us and our co-workers (Draper *et al.*, 1993) for insisting that some 'leap of faith' is always present in prediction, and for providing only a circular argument. But their arguments in support of routine predictability also seem circular: a 'law' holds except when it does not. Our differences seem to be merely semantic. Their formulation in the third paragraph of Section 6.3 of the 'basic statement of empirically based induction' can be stated in our language as

> 'If past observations are judged exchangeable, and the conditions under which they were observed are judged exchangeable with a new set of conditions, then we can predict that the event will recur in the new situation'.

We can predict, but we may be wrong; we need to have faith that no new factor is influencing the outcome. Some sort of leap of faith about the persistence of structure (Hodges, 1987) is always present in prediction.

Stephen G. Hall (London Business School): This paper contains much with which I agree. If we take its main point to be a call for the analysis of all available data then few could argue with it. If in addition it is arguing for a depth of insight which can emerge from using data from different time periods then my own work and recent publications on co-integration would support the conclusions drawn here.

I believe that the data under consideration are characterized by two features. First, it is a panel of data consisting of a number of cross-sections taken at different points in time. Second, because the moments of the distribution of the cross-sections change over time they are drawn from a non-stationary population. This last point is crucial to the procedure proposed by the authors as otherwise the sets of data would only differ by sampling error and so the line drawn through the means of the data sets would be meaningless. Much recent work has been undertaken on appropriate estimation and inference strategies for models involving non-stationary data. Although the procedure proposed in this paper will clearly give consistent estimates of a well-specified model, my concern is that in the absence of any theory of inference the rejection of incorrect models becomes a matter of arbitrary interpretation on the part of the researcher. Also, although not so relevant to the types of data set used by the authors, the technique proposed here is an inefficient way of using the data for small data sets.

The theory of non-stationary regression offers a formal framework for testing models of this type. It also allows the treatment of more complex models, both in terms of the number of variables and the dynamic structure of the model, and a better understanding of the efficiency considerations when data sets are small. This paper points to some serious weaknesses in traditional regression theory and this I fully applaud, but this does not mean that the whole body of statistical theory should be abandoned. The theory can, and has been, extended to account for these problems and I believe that it offers a more comprehensive way forwards than the procedures proposed here.

D. V. Lindley (Minehead): The main idea in this paper that we should often be looking for general rules, rather than models for single situations, is brilliant. The authors are correct to criticize statisticians for the comparative neglect of the topic. I would like to add a more formal structure to their argument, using subjective probability.

For several data sets, let X_i denote the data in set i ($i=1, \ldots, n$). It is usual to model this by a probability distribution $p(X_i|\theta_i)$ dependent on a parameter θ_i. It is then necessary to assign a distribution for the parameter set $\boldsymbol{\theta}=(\theta_1, \ldots, \theta_n)$. Let this be $p(\boldsymbol{\theta}|\phi)$, where ϕ is a hyperparameter. One possibility is to suppose the parameters to be independent and identically distributed given the hyperparameter. This general model is termed hierarchical. The main interest may then be in the hyperparameter, expressing the general rule that is being investigated, whereas the parameters express the behaviour in narrower cases. With a distribution for ϕ, the model is complete and other probabilities can be evaluated. For example, general prediction of Y from the data is provided by $p(Y|\mathbf{X})$. The evaluation of this would depend on whether Y was from one of the previous sets or from a new set. Details are standard and omitted.

The advantage of this type of procedure is that it gives an explicit way of calculating all quantities of interest within the framework of specific assumptions that are spelt out. The idea can easily be extended to include nuisance parameters at each stage of the hierarchy. Additional stages can be included if needed. Computation can sometimes be difficult, but progress is rapid here and even fairly complicated models can be handled.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

R. Murray Lindsay (University of Saskatchewan, Saskatoon): This paper will make anyone interested in methodology pause for self-reflection about the validity of our current research tradition and conception of knowledge. In presenting a more realistic picture of science, where progress requires laborious work, extensive insight about the subject, and much judgment, the authors highlight the folly of excessive reliance on 'yes–no' type of decision-making based on the outcome of significance tests. It is inappropriate to base inferences as if each study was conducted in isolation, without the benefit (or possibly harm!) of incorporating previous knowledge. More importantly, a concrete operationalization of the elusive notion of 'significant sameness' is explicated, whereby the criterion of adequacy is whether the *same model* holds over *many* sets of data. Meta-analysts take note!

The paper reinforces the need to appreciate Deming's distinction between analytical and enumerative studies. A basic assumption underlying mainstream research is to collect data focused on the discovery of statistically rigorous, universal relationships by using large samples. However, this only produces aggregate generalizations, say that some parameter differs in the overall (super)population. We cannot infer that the result or theory applies to every member, which is the essence of analytical generalization and, according to the authors, scientific practice. This requires examining specific conditions to see whether the result changes.

Finally, the authors indicate that obtaining routine predictability of 'stubborn facts' early on is necessary to provide a dependable basis for theorization. Despite the 'testing' machismo of our vernacular, likely to assist in getting results published, the underlying reality is that much research in the social and behavioural sciences is exploratory, whereby the aim should be to obtain reliable facts, to determine the population(s) to which they apply, and to begin to inquire about the processes that might underlie the results. To continue to operate under false pretences will ensure that such disciplines remain immature.

I have one small criticism. Although there is recognition that there is no trans-historical, neutral, permanent language (criteria) for evaluating both theories and observations, the authors fail to acknowledge the merit of other research approaches. For example, the interpretive perspective sees social reality as being emergent, subjectively created and objectified through human interaction, often resulting in unstable definitions and responses (Chua, 1986). In this research tradition the aim is not to develop social prediction rules but rather to understand and interpret social action (Blumer, 1969). Attempting to focus exclusively on the former could prevent the development of new and valuable insights.

In conclusion, this is an important paper which can be expected to contribute significantly to our understanding of scientific practice.

John W. Tukey (Princeton University): As usual, I agree with Andrew Ehrenberg somewhat more than I disagree. Emphasis on the importance of a more nearly correct functional model—rather than on a more nearly correct stochastic model—is something that all statisticians should accept—for a variety of reasons, including those put forward by Ehrenberg. Since it is easier to verify the functional model (how many real detailed verifications of stochastic models does any one of us know?), we need to emphasize combinations of more nearly correct functional models, with forms of analysis that are robust in not requiring the stochastic model to be at all precisely known.

Granted that new simple fundamental relationships are of the highest importance, we must still recognize that no more than one physical or chemical experiment in 1000 involves anything like a new simple fundamental relationship. I see no reason why the ratio in 'social science' fields like marketing should be greatly different.

We need to recognize that Ehrenberg's argument for

$$w(1-b) = \text{constant} \quad \textit{versus} \quad w = \text{constant}^* + (\text{constant}^{**})\,b$$

is a conventional instance of parsimony—one constant per situation instead of two. This is important—but not as far from conventional analyses as we might have thought. Many aspects of conventional analysis complement Ehrenbergian analysis, and certain of them are likely to be important. Thus, the authors' Table 7 is, by itself, terribly weak support for $w(1-b) = \text{constant}$. We can believe that the observed ws might be a monotonic function of b, but even this is at best hazy (and quite disheartening). Asking how good the fit seems to be is well worth our attention. (The law of DuLong and Petit is as simple as $PV = \text{constant}$, but its lack of precision has led to its being nearly forgotten.)

This of course leads us to the important point, underplayed by Ehrenberg and Bound, that the next step after finding a simple model is to study the deviations of practice from simple theory. In Table 7, for example, should we be looking at the ratio '(observed w)/(predicted w)' and relating this to the

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

geographic distributions of the stores of each chain in the light of demographic differences in customer families?

The **authors** replied later, in writing, as follows.

Several of the discussants claimed that predictive success must depend on whether one is dealing with the physical or social sciences, or with experimental or observational data, or deterministic or stochastic variables, or static or dynamic situations, or large or small samples, or whatever. We disagree. Our two examples straddled all these dichotomies, as did our discussion.

Many of the speakers also referred to predictions such as are made in forecasting or policy decisions which we had explicitly excluded from the start. Our stated concern in this paper was solely with inductive predictions as in ordinary science (with a small 's'), based on many sets of data (MSOD). David Cox therefore rightly notes the importance of studying relations under a known range of different conditions. But he is far too complacent, if not plain wrong, when he says that this has already been stressed in the statistical literature. A reference to one obscure paper 55 years ago and our two equally obscure ones is not enough. Why did he and the other famous authors in question not talk about it elsewhere?

Several discussants asked how one defines MSOD. Table 1 is based on five sets of data, Tables 2 and 3 give 15 more, and Table 6 refers to yet more. One cannot miss MSOD when one sees them (if one is prepared to look). More formally:

'Data consist of two or more sets when observing a reading in one is operationally independent of observing a reading in the other'.

In contrast, a single set of data (SSOD) is a set of undifferentiated readings, e.g. a sample from the UK population. If one divides this into men and women, rich and poor, north of Watford or south, one begins to have MSOD. However, statisticians usually pool any MSOD that they happen to come across so that they can apply their SSOD techniques, such as least squares or Lindley's probabilities. If Chris Chatfield insists on never going beyond an SSOD he will not experience the joys of successful prediction.

Peter Urbach, as a philosopher of science, also treats his observations as an SSOD. Only *very* bad scientists would be as unobservant as Urbach and Hume—the only thing that they report about their 17 white swans is that they were 17 white swans, an undifferentiated SSOD. If Urbach had noted (or ensured by design) that some came from China, and others from France and Brazil, and that some were young and some old, male or female, big or small, observed in the winter or after moulting in the summer (if swans moult), he could have predicted with near certainty that within that range of conditions his 18th swan would be white too.

We have never in fact said that we saw Hume's problem of induction—about any number of such identical observations—as central. Instead we just thought he had asked the wrong question. Our resolution of it in Section 6.3 should presumably have been phrased yet more emphatically:

'Given successful replications over a stated range of *different* conditions, one can predict . . .'.

Urbach has in fact understood this sufficiently well to query what conditions would be relevant. *A priori* the answer is ALL. As Professor Davies notes, all analysis situations involve many variables. The skill is to pick on situations where most of these variables do not affect the result, i.e. have zero slope coefficients. Paradoxically, these 'irrelevant' variables are then the very ones which are particularly relevant, because they give predictive meaning to the result: that it continues to hold despite that these variables vary. One can therefore predict that double jeopardy (DJ) will hold for Australia if it has already held there, so that its being Australia has been found to be irrelevant. But one would not want to predict that swans would be white there—this is not because Australia is different, but because the swans there were.

Draper and Mallows will also continue to be locked into their vicious doom loop of having to perform 'leaps of faith' if they still will not recognize that exceptions do not just occur when they occur but can be pinned down (black swans occur in Australia and are no longer worth a letter to *Nature*).

The predictive question remains whether the result has already remained the same for different data sets. Sidney Rosenbaum refers to people's heights and weights which are correlated. But was the relationship the same in his different data sets? Using Quetelet's index, he reports coefficients all of which differed. The model is therefore not predictive (just like our first example in Section 2.1), although Dr Rosenbaum fails to recognize this failure. Robert Fildes rightly refers to models with grandiose claims but no empirical basis (like Professor Davies's 'axioms'?) which then not surprisingly fail.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

Chris Chatfield asks for cookbook guidance to find the functional form of a predictive model (but he oddly sees this as a task only for scientists and not for statisticians like himself). The basic criterion is that the same function has to have held across MSOD. That is enough. It produces the traditional simplicity of good science. That is why the one-parameter hyperbolas $w(1-b)=w_0$ score so much over John Copas's arrays of different two-parameter straight lines; and similarly for Boyle's $PV=C$ hyperbolas. But to find suitable functional forms like this (i.e. ones which generalize) takes knowledgeable and creative cooks, not cookbooks.

The single parameters w_0 or C as such do not of course cause any problems: they are determined, in effect, by putting the curves through the means of the two variables. In DJ, w_0 reflects the fact that the base-line average purchase frequency differs from case to case, e.g. with the product category (e.g. cereals or cooking oil), with the length of the analysis period, and with the type of consumer (e.g. by country, or large *versus* small households). In Boyle's law, the C reflects that the absolute levels of pressure and volume differ from case to case with the amount and type of gas and its (controlled) temperature.

We were somewhat surprised when John Copas did not see that the only real problem in fitting straight lines is with the second parameter, the slope coefficient. We had thought that every statistical schoolchild would know this, at least implicitly. But now that we think of it, we have never seen the point discussed like that in the literature.

What matters ultimately is understanding as well as brute empiricism. Here David Cox has it totally wrong when he says that we massively undervalue the role of theory. Firstly, our paper is ALL about using prior knowledge, i.e. analysing data when one already knows the relationships such as $w(1-b)\simeq w_0$ or $PV\simeq C$. This is (low level) subject-matter theory. As John Tukey so rightly says, no more than 1 in 1000 experiments in the natural sciences involves a *new* relationship. Yet classical statistics—of which we are very critical—provides only methods for supposedly discovering a new relationship.

Cox also ignores that our relationships such as $w(1-b)=w_0$ and the Dirichlet model were derived by theoretical arguments from deeper assumptions. John Copas here is unpardonably simplistic, for a statistician, in querying one of these. He somehow doubts that the buying of Rice Krispies and Corn Flakes can be effectively independent, as assumed. But the correlation is always nearly 0 (or more precisely about 0.1—see Ehrenberg (1972) and Goodhardt *et al.* (1984)). Independence only means that knowing whether or not people bought Rice Krispies in a year, say, does not help one to predict whether or not they also buy Corn Flakes in that analysis period. What is so odd about that?

Theory must in fact ultimately be well grounded. Our paper is crucially about ways of dealing with the MSOD involved. A recent comment on Darwin in the *London Review of Books* lauded his 'brilliant talent for intuitive speculation', just the sort of thing that David Cox so rightly likes. But the reviewer went on to say that 'this meant nothing (!) . . . until supported by the infinitely painstaking empirical researches, ceaselessly extended, which occupied his working days'. David (and others) please note: statisticians are on the whole not prepared to analyse the MSOD that are involved—hence their collective predictive failure, and our paper.

Several discussants queried our assertion that there was in fact no success story in the history of regression analysis. David Cox claims that only lack of space prevented him from discussing examples. The rules of the Society allowed him almost 200 more words, enough for several dozen references together with his typically succinct outlines of two or three such cases. Some discussants even asked for a prize. At the meeting we therefore renewed an earlier offering of a prize, now standing at $10000. (The formal conditions will be set out by one of us (ASCE) in the Society's newsletter *RSS News*.

In conclusion, John Tukey worries about our Table 7, which had been Professor Keng's first look at DJ for people's *store choice*. This provides an excellent base to illustrate further the 'revolutionary' MSOD doctrine which we are pursuing here (as long practised in the physical sciences).

The correlation between the observed w and predicted $\hat{w}=2.6(1-b)$ in Table 7 is only 0.71. John finds this 'terribly weak', 'at best hazy' and 'quite disheartening'. (What does he make of the rest of social science?)

But what happens next? John Tukey suggests studying the deviations. Well, they were irregular, and with no outliers: -0.2, 0.6, -0.2, 0.6, -0.6. That is 'good'. But it does not tell us very much more.

We do, however, know a little more about the deviations because of their MSOD background (i.e. some of the 'excluded variables'). This tells us that the Coop is a co-operative, Kwiksave a discount chain, Tesco large nationally, Asda carried no private labels at the time and Fine Fare was

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms

small. Does this help with the deviations?: not much, by itself. What did John Tukey have in mind for 'studying the deviations'?

The approach outlined in our paper is about something totally different, namely establishing whether the DJ relationship $w(1-b)=w_0$ holds also for *other*, *different*, *data*. This has many advantages. If DJ does not recur, one need no longer worry about the deviations in Table 7: it does not represent a generalizable relationship anyway. But as we said in the paper 'DJ for store choice has also been found for other products and countries and with higher correlations . . .'. One therefore now has a generalizable predictable result, within a much wider empirically based context and with explanatory theory to boot. That is good science. One still need not worry unduly about the deviations in Table 7 (having seen that they are fairly small and irregular).

Professor Davies finds all such things potentially 'banal'. (*Bad* examples of undiluted empiricism no doubt are.) It is supposed to lack in excitement and challenge, and also that 'little statistical input is required'! The real difference, however, is, we suspect, whether one prefers continually to wallow unsuccessfully in messy looking but 'promising' data, or to establish the simple and routinely predictable relationships of normal science. We would happily offer many token $1 prizes for cases where predictive models based on MSOD have been successful with new but different data sets.

REFERENCES IN THE DISCUSSION

Blumer, H. (1969) *Symbolic Interactionism*. Englewood Cliffs: Prentice-Hall.

Bound, J. A. and Ehrenberg, A. S. C. (1989) Significant sameness. *J. R. Statist. Soc.* A, **152**, 241–247.

Chua, W. F. (1986) Radical developments in accounting thought. *Acctng Rev.*, Oct., 601–632.

Colombo, R. A., Sabavala, D. G. and Ehrenberg, A. S. C. (1993) The car challenge. *Working Paper*. New York University. New York.

Draper, D., Hodges, J. S., Mallows, C. L. and Pregibon, D. (1993) Exchangeability and data analysis (with discussion). *J. R. Statist. Soc.* A, **156**, 9–37.

Ehrenberg, A. S. C. (1972) *Repeat-buying*. London: Griffin.

Fildes, R. and Makridakis, S. (1993) The impact of empirical accuracy studies on time series analysis and forecasting. *Working Paper OR 1/93*. Management School, Lancaster University, Lancaster.

Floud, R. C., Wachter, K. W. and Gregory, A. (1990) *Height, Health and History*. Cambridge: Cambridge University Press.

Goodhardt, G. J., Ehrenberg, A. S. C. and Chatfield, C. (1984) The Dirichlet: a comprehensive model of buying behaviour (with discussion). *J. R. Statist. Soc.* A, **147**, 621–655.

Hodges, J. S. (1987) Uncertainty, policy analysis, and statistics (with discussion). *Statist. Sci.*, **2**, 259–291.

Hulme, C., Thompson, N., Muir, C. and Lawrence, A. (1984) Speech rate and the development of short-term memory span. *J. Exptl Chld Psychol.*, **38**, 241–253.

Jordan, T. E. (1993) *The Degeneracy Crisis and Victorian Youth*. Stony Brook: State University of New York.

Jung, C. G. (1972) *Synchronicity: an Acausal Connecting Principle*. London: Routledge and Kegan Paul.

Kauffman, S. A. (1991) Antichaos and adaptation. *Scient. Am.*, **265**, no. 2, 64–70.

Knight, I. (1984) *The Heights and Weights of Adults in Great Britain*. London: Her Majesty's Stationery Office.

McGinnis, R. (1968) A stochastic model of social mobility. *Am. Sociol. Rev.*, **33**, 712–722.

Rosenbaum, S. and Crowdy, J. P. (1992) British Army recruits: 100 years of heights and weights. *J. R. Army Med. Corps*, **138**, 81–86.

Ward, D. (1987) *The Validity of the Reconviction Prediction Score*. London: Her Majesty's Stationery Office.

Yates, F. and Cochran, W. G. (1938) The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556–580.

This content downloaded from 90.253.172.169 on Thu, 10 Oct 2019 04:19:29 UTC
All use subject to https://about.jstor.org/terms