/ RICHARDSON

A full-scale portrait
lgrimage, the serial
as a pioneer of liter-
d a masterwork of
iction. The book
ment of *Pilgrimage*,
f how and why it was
ent to which, as the
) be written and
lson's life and art
ges. $15.00.

rbana, Illinois 61801

the American Scholar

Published for
General Circulation
by
PHI BETA KAPPA

Advertising Representative: KERR & ASSOCIATES, 1211 Connecticut Avenue, N.W., Suite 604 Washington, D.C. 20036. Telephone: 202-296-1478.

Published quarterly since 1932 by the United Chapters of Phi Beta Kappa, 1811 Q Street, N.W. Washington, D.C. 20009. Second-class postage paid at Washington, D.C. and at additional mailing offices.

Send all change of address notices, undeliverable copies, subscription orders and other mail to *Editorial and Circulation Offices:* 1811 Q Street, N.W., Washington, D.C. 20009.

Subscription rates: $2.00 a copy, $8.00 one year, $14.50 two years, $20.00 three years, foreign and Canadian $.75 a year additional for postage.

Unsolicited manuscripts must be accompanied by return postage. No responsibility is assumed for their loss or injury, or for the views expressed in articles published.

Formulas, Numbers, Words: Statistics in Prose

WILLIAM KRUSKAL

Statisticians deal professionally with symbols. Ultimately our symbols stand for, or relate to, sacks of potatoes, political opinions, illnesses of people—that is, to relatively material, palpable objects, to actual things that should always be kept in mind. Yet in abstraction and abstruseness, statistical discussions, like those of any scientific practice, may be far away from potatoes, presidential nominees, and peritonitis morbidity.

The symbols we use are, in the first place, numbers, but numbers alone are insufficient. For sensible statistical treatment they require some kind of theoretical, mathematical coherence, however simple and however tacit it may be. Thus even to add two different numbers of gallons of oil presumes disjointness of objects, the same unit of measurement, and an additive scale. In even slightly more complex circumstances, sensible treatment of the numbers may rest on a complex set of assumptions about statistical independence, approximate normality of distribution, and other properties of considerable abstractness.

So, starting with numbers, we are inexorably led to the abstract, mathematical side of statistics—call it, for brevity and at the risk of distortion, the side of *formulas*. It includes formulas, taken literally, together with much mathematics from the simple to the deep, and a range of philosophical thought as well.

There is a third side of statistics that corresponds to a third kind of symbol: the side of *words*. Numbers, formulas, and words make up a useful trichotomy, and here I shall primarily discuss words in statistical theory and practice. Important as they are, words get short shrift in most statistical teaching and exposition, even though these are done, to a large extent, in terms of words. Perhaps we statisticians feel generally that the study of words belongs to the psychologist, the linguist, the scholar of language, but not to the statistician as such (unless, of course, the statistician is helping the psychologist or linguist with his empirical studies).

Yet words play a fundamental role in all statistical activity. The most abstruse mathematical statistics requires words for motivation, explanation, and perhaps even creation. Expositions of numerical statistical analyses require words, and different words for different audiences; I might, looking ahead, even write "significantly different words." Documentation of numerical data and documentation of statistical computation procedures require words, and the producers and consumers of such documentation recognize that to provide clear wording is a far from trivial task.

Perhaps one reason for statisticians' widespread indifference to words is that we are determined, along with Humpty-Dumpty, to be masters of the symbols we use. As Humpty said, "When *I* use a word . . . it means just what I choose it to mean—neither more nor less." Yet Lewis Carroll himself was probably sensitive to the impossibility of such a solipsistic program. First, there would arise the infinite-regress problem of defining words or symbols more generally in terms of still other symbols. Second, to use words in wholly individualistic ways is to invite inattention, needless controversy, and regressive confusion. Even for purely mathematical symbols, indi-

WILLIAM KRUSKAL is Ernest DeWitt Burton Distinguished Service Professor of Statistics at the University of Chicago, where he is also currently serving as dean of the Division of the Social Sciences.

vidualism to an extreme would be troublesome: suppose someone began to use multiplication signs to denote addition. A still simpler case is the mild confusion that sometimes results from the different conventions used in America and Europe for the decimal point and the comma.

(Not everyone would agree with my suggestion that Lewis Carroll intended as ironic his often cited Humpty-Dumpty statement. For example, Martin Gardner, in *The Annotated Alice,* interprets Carroll literally. Yet Gardner himself is well aware of the dangers in symbolic eccentricity. His theological novel, *The Flight of Peter Fromm,* contains the following apt passage: "Shall I prove to you that all visible objects are green? It is easy. I define 'green' as the property of reflecting light. All visible objects reflect light, therefore all visible objects are green. For the impregnability of such assertions, however, I must pay a heavy price. Not only have my statements become vacuous, but because my language departs so widely from common speech my assertions are sure to cause more confusion than enlightenment. Of course I then can have the pleasure of explaining to my critics how they misinterpret me.")

Thomas Hobbes had a theme somewhat like Humpty-Dumpty's when he wrote in *Leviathan* that "words are wise men's counters, they do but reckon by them; but they are the money of fools." Yet no matter how much words are but counters, to be arbitrarily assigned, those assignments require considerable consistency and sharpness.

A common misuse of statistical words lies in the transplantation of reasonably precise technical terms into inappropriate contexts. The term "statistically significant" is a frequent victim. In a recent letter to the *New York Times,* for example, a letter about our postal system, the writer says:

> The vastness of this country, the high mobility rate of many of its inhabitants and its statistically significant immigrant population all contribute to the need for an efficient postal service.

That "statistically significant" gives the sentence a lot of rhetorical weight, although all it appears to mean is "large." A similar case appears in a *New York Times* article by Harold M. Schmeck: " . . . the numbers involved in this comparison were considered too small to be statistically significant."

This difficulty is not limited to the ephemeral columns of newspapers; one may find it in highly regarded scientific journals. For example, in a 1975 issue of *Science,* the official journal of the American Association for the Advancement of Science, an article on temperature trends says that

> It is difficult to test our results . . . against observations because no statistically significant *global* record of temperature back to 1600 has been constructed.

Here the significant phrase apparently means "accurate," but I am not certain.

A 1973 example from the same journal says that

> . . . if we take n pictures of N_0 particles, . . . [we have] $6nN_0$ coordinates, a number which can easily be made sufficiently large to have statistical significance.

Here again the term is used in a vague, almost boastful sense.

In fact the phrase "statistical significance" has a widely understood, rather precise meaning that is taught in introductory statistics textbooks and courses. It refers to a so-called tail probability that is surprisingly small under a relevant, if sometimes tacit, hypothesis. For example, the hypothesis may be that a new soporific is no better than a related standard one. The experiment may use as subjects ten pairs of people, perhaps matched for age, sex, state of health, and the like. One randomly chosen member of each pair uses the new soporific; the other member uses the old. If in all ten pairs the new soporific does better, we might well say that the new is better than the old with statistical significance—in fact, at a so-called significance level of $1/2^{10} = 1/1{,}024$, or a bit less than 0.001. That small number represents the probability that the superiority of the new soporific would be observed in all ten pairs *if* the hypothesis of indifference really were true. Suppose, next, that the new soporific were seen to be superior in just *nine* of the ten cases. One would still be likely to write of statistical significance, this time at the significance level $(10 + 1)/1{,}024$, or about 0.01. That number is the probability, under the indifference hypothe-

sis, of finding what in fact we did find (nine out of ten) or something more extreme (ten out of ten). But enough; this is an essay about statistics, not a lesson in statistics.

Another often misused statistical expression with a precise technical meaning is "random sample." I shall not present a one-paragraph introductory description here (although that would not be difficult), but shall content myself with illustrating the range of misuse.

One misuse is pejorative; for it, a random sample is an erratic grab bag. An example of that misuse occurs in a book review by Allan K. Wildman in the *Journal of Modern History:* "What we are given is a careless and incomplete random sampling to illustrate a passing point." At the other extreme, the term "random sample" may be used to provide a spuriously positive scientific effect, like the white jacket worn by an actor in a television toothpaste commercial. For example, in a 1973 *New York Times* article, Lawrence K. Altman reports a Swedish study on views of death. He paraphrases a Swedish cardiologist: "Dr. Biörk . . . emphasized that his analysis was . . . not a scientific study based on a random sample." In terms of the statistical meaning of "random sampling," it is neither necessary nor sufficient for ensuring that a study be scientific.

There are also expressions that do *not* have precise, commonly accepted statistical meanings, but are widely (and wildly) misused in a profusion of senses in both the lay and the scientific literature. My own favorite is the term "representative sample." That term, with its built-in begging of the question, has an authoritative ring that forms an orotund trap of fake eloquence and fake precision.

Sometimes the term "representative sample" is used solely as decoration. A nice example is provided in a 1975 *New York Times* column by Raymond J. Cormier that appeared twice, first in French and then in English. Cormier wrote first that ". . . dans un échantillon de 400 universités représentatives, les inscriptions aux cours du français ont baissé de 30 pour cent." Next day the English version appeared: ". . . in a nationally reliable sample of 400 colleges, French enrollment went down nearly 30 percent." Neither the idea of a representative university nor of a nationally reliable sample has an accepted, sharp meaning in statistics.

Sometimes the representative idea is wholly confused in what appears at first to be a technical statement. For example, an article on political polls by Ken Bode in the January 17, 1976, *New Republic* quotes an unnamed pollster: "Every time you draw a sample you have a 95 percent chance that it accurately represents the total population and a five percent chance that it does not."

Another sense is that of diversity. The 1975 Annual Report of the Western Publishing Company says that the "representative sampling of products shown on the following pages graphically illustrates the diversification and appeal of our lines." And sometimes "representative" means typical; it is used that way by Virginia Woolf in *Three Guineas,* where she says, after quotations from the Marquess of Londonderry, Winston Churchill, and a Mr. Cyril Chaventry of North Wembley: "That opinion, too, is a representative opinion, one from a great number to the same effect."

On the other hand, Ralph Waldo Emerson, in his *Representative Men,* meant by "representative" the ideal quintessence of a type. Plato is the representative philosopher, Goethe the representative writer, and so on.

The blur between typical and extraordinary is at first puzzling, but C. S. Lewis gives an explanation in his *Studies in Words:*

> . . . this demand for the typical easily merges into a demand for the perfect. The quest for the wholly normal cabbage . . . would involve the rejection of every cabbage which had suffered from such historical accidents as bad soil, unequal sun . . . , too much or too little rain, and so on. In the end you would be looking for the ideal cabbage.

It is a temptation to continue with these problems of representativeness, but I move on to a related example, the use of the word "explain" in expositions of the concept of correlation, or in correlational analyses of data. We see statements like "Hours of study explain 62 percent of the variance in test scores." The danger here is that "explain" comes perilously and ambiguously close to "cause." As Edward Tufte says, in his *Data Analysis for Politics and Policy:*

> . . . some researchers . . . play on the ambiguity of the word "explain" . . . to avoid the risk of making

an out-and-out assertion of causality while creating the appearance that something really was explained substantively as well as statistically.

Another generic way in which choice of wording affects atmosphere arises in trials of a new medical treatment. An enthusiastic reporter may write that the "new treatment is better than the conventional treatment with statistical significance at the *0.05* level." A more neutral writer might say that "the 95 percent confidence interval for the new treatment's cure rate, minus that of the old, is from 1 percent to 11 percent." The second statement formally implies the first, yet what a difference there is in the impression left in the reader's mind!

As we see, the statistical field of hypothesis testing is a rich source of verbal misuse. Some years ago, I had many conversations with the late L. J. Savage about his criticisms of hypothesis testing as ordinarily exposited and practiced. I agreed with many of those criticisms, although they did not all apply to the core ideas sensibly interpreted, but rather to the way those ideas had become inflexible, conventional, and vulgarized over time. Savage was not pleased when I predicted that his own subtle, creative, and ingenious ideas about subjective probability would, if they caught on, inevitably become rigidified and vulgarized; yet that is what has happened. Such a process seems to be inevitable: success breeds vulgarization. Lionel Trilling in his book *The Liberal Imagination* puts it thus:

> . . . organization means delegation, and agencies, and bureaus, and technicians . . . the ideas that can survive delegation, that can be passed on to agencies and bureaus and technicians, incline to be ideas of . . . a certain simplicity; they give up . . . largeness and modulation and complexity in order to survive.

We have already touched on the misuse of the idea of statistical significance. Two other paired terms are the mysterious ones of "acceptance" or "rejection" of a hypothesis. Endless argument has been based on overly literal use of those words, which truly are Hobbesian counters but are misunderstood because they are redolent of everyday psychological and commercial interpretations. In technical statistics, "acceptance" and "rejection" of a hypothesis are simply canonical words for two actions or decisions, one of which would be preferred were the hypothesis in fact true, and the other if it were in fact false. We may look down our noses at the oversimplification represented by the dichotomies; but, given those dichotomies, it is good to have words for the moieties. It is misleading, however, to have words with such emotional connotations.

I am not tilting at windmills here. One of the most popular introductory texts for statistics falls squarely into the connotational trap; it argues that we never *really* accept the null hypothesis, we just fail to reject it. Hence we need not worry about error of the second kind, or power. The author of the text even continued that fraudulent message from his first edition through later ones, thus marking another triumph of stubbornness.

Yet to use denatured, flavorless words would deprive us of liveliness and retard the study and spread of new ideas. Suppose, for example, that catastrophe theory (a recent mathematical development) were called the "theory of discontinuities from folded surfaces." Or, to turn to a central area of statistical theory, suppose that Sir R. A. Fisher—a master at public relations—had not taken over from ordinary English such evocative words as "sufficient," "efficient," and "consistent" and made them into precisely defined terms of statistical theory. He might, after all, have used utterly dull terms for those properties of estimators, calling them characteristics *A*, *B*, and *C*—or, more egocentrically, *R*, *A*, and *F*. Would his work have had the same smashing influence that it did? I think not, or at least not as rapidly. Still, the scope of the concepts can certainly be questioned. Sufficiency may be altogether spoiled, for example, by tiny changes in the distributions specified by the underlying model; so may efficiency and consistency.

Or turn to Fisher's wonderful phrase "analysis of variance," and the accompanying conventional tables in which columns add up so neatly. Is it too cynical to think that the lovely term—half-mystery, half-promise—and the orderly tables helped to win acceptance, quite aside from the underlying theory?

There is the dilemma. We want to be objective and we value the trappings of objectivity—for example, somber tables of numbers and terrifying pages of formulas. Yet

when we write in words about numbers and formulas we seek acceptance and readership, so we are moved to make up exciting terms, and exciting terms may mislead. In 1938, Thomas Mann wrote in *The Coming Victory of Democracy* that "humanity, in small things as in great, strives for variety, changes, for the new, because it promises him an amelioration and an alleviation of his eternally semi-painful condition." If R. A. Fisher chooses to clothe his great ideas in exciting words, one cannot fiercely object. The nub of the difficulty is presented by the false Fishers, whose exciting presentations cover triteness or even nonsense.

Such issues must be faced by all of us in teaching, talking, and writing. As I prepare this essay, I ask myself the unanswerable question whether the anecdotes and quotations I use to try to keep your attention may not be shameful and meretricious. The only basic response, of course, is discussion and more discussion. So I hope that my remarks—shameful and meretricious though they may be—will stimulate some of you to discuss these difficulties with the use of words at a depth you do not ordinarily plumb.

In *The Small House at Allington*, Anthony Trollope, writing about his Plantagenet Palliser—later to become prime minister—as a young man, notes: "Statistics were becoming dry to him, and love was very sweet. Statistics, he thought, might be made as enchanting as ever, if only they could be mingled with love."

I digress for the sake of love to a brief discussion of one of the more fascinating words in statistics, the word "normal." We have in statistics the normal distribution and the normal equations. In other parts of life there are normal schools, normal hydrocarbons, normal ranges in medicine, and (some time ago) "back to normalcy" as a political slogan. The C. S. Lewis passage quoted earlier mentioned the normal cabbage.

To call the normal distribution normal is more than an arbitrary naming; a normative aspect is one that persists. Some statistics textbooks published even in this decade misleadingly say that the normal distribution is so called because "it is the distribution normally encountered."

The origin of the phrase "normal distribution" is unclear. Stephen Stigler, professor of statistics at the University of Wisconsin, and I have been trying to track it down, with partial success, by starting with the writings of Sir Francis Galton. He wrote *about* the normal distribution at great length, and with an amazing range of elegant variation. Among the terms he used were "the exponential distribution," "the law of error," "the law of deviation," "the Gaussian Law," "the law of statistical constancy," and, of course, "the normal law," the last first used by Galton in 1877 as far as we can find. Karl Pearson claimed at a later time that he himself had coined the term "normal" as a neutral word to avoid the chauvinistic, competing, nationalistic "Gaussian" and "Laplacian," but it seems to us that his memory was wrong.

Our current hunch is based on the observation that Galton and his contemporaries, when writing of the normal distribution, not only used terms like "the exponential distribution" but, to avoid monotony, brought in vaguer expressions: "the usual distribution," "the commonly encountered distribution," and, of course, "the normal distribution." (Indeed, "normal" in that sense was used still earlier, in 1873, by Charles Sanders Peirce, the great American philosopher and statistician.) We think that somehow "normal" won out among the synonyms. Why should it have won? I speculate that "normal" is a word with powerful, positive connotations because of the ambiguity between its two meanings: (1) something desirable and (2) something commonly found. A major theme in our culture, after all, is the desirability of what is commonly found, so the two senses reinforce each other.

The history of the term "normal school" is related to that speculation. Why should a teacher training institute be called normal? Presumably because it is a model, but a model for what? For other teacher training institutes? When Stigler and I began to look into this question, we were told that the term went back to the founding of the École Normale by Lakanal in 1794 during the French Revolution. Yet we quickly found that the term went back at least to usage in Vienna in 1770 by an ambitious pedagogue named Joseph Messmer. As part of Maria Theresa's educational reforms, Messmer soon became head of the first *Normalschule*, a school for youngsters to serve as a model or norm. To it

was attached a teacher training establishment that used the school for practice and observation by student teachers. As time went on, the word "normal" apparently shifted from the model school for kids to the school for teachers. After all, teachers are older, bigger, and tougher; is it surprising that they should appropriate the best adjective? Soon the term came to be applied to teachers' schools even without children.

I turn next to difficulties of wording in the public documents of government statistical agencies. Those agencies have special problems with maintaining objectivity, in part because of pressures—at any rate, feared pressures, or alleged pressures—stemming from partisan politics. In addition, government documents issue as a rule from agencies, not individuals, and their official status requires at least a strong attempt at impersonality.

One need not be a naïve Marxist or a cynical marketing expert to agree that cultural predispositions affect government statistics at fundamental levels. Even such concepts as the household, head of household, address, marital status, race, and so on are reflections of our ambient culture, and they lead to difficulties as that culture creaks, groans, and changes. Individuals without fixed addresses or attachment to specific households, for example, may cause irritating definitional problems. Matters get much worse for such studies as crime victimization surveys, in which respondents are asked about crimes against themselves or their families. What one person may regard as a forceful attack may not be so regarded by another, and if that difference is confounded with other characteristics (of income, education, and so on), fresh problems of interpretation arise.

Transition from a table of numbers, perhaps the results of a survey, to a statement in words is rarely simple. The monthly employment tabulations of the Bureau of Labor Statistics, the monthly consumer price index figures, the periodic tabulations of crime frequencies, imports and exports, school registrations, and so forth are typically accompanied by statements in words of what they mean. How is this to be done? What are the important numbers, the major shifts from last month or last year?

Here enters a standard ambiguity. Should importance of changes be measured statistically—say by statistical significance via multiples of standard errors of differences—or substantively according to their economic, criminological, or other import? Perhaps by both. A basic trouble with substantive import is that it is relatively more subjective, and I suspect that in many cases statistical significance is the primary requirement for mention of a change as part of a routine transition from numbers into words. Of course this can get one into classic troubles. For example, it might happen that both white and nonwhite unemployment shift *less* than 1.6 times their standard errors (or whatever the conventional multiple is), yet total unemployment shifts *more* than 1.6 times its standard error.

In the early seventies, the Bureau of Labor Statistics used the term "significant," or at least "non-significant," in its releases, without making it clear whether statistical or substantive significance was intended. Recently the bureau seems to have dropped the word "significance," usually contenting itself with flat statements, sometimes qualified by modifiers like "slightly" or "relatively small." That hardly settles the issue, for there have to be policies about when a level or a shift should be mentioned, and in what terms.

The problem for the agency is difficult. On the one hand, it can hardly be guided by a fixed algorithm in going from numbers to words; on the other hand, any ad-lib human intervention creates subjectivity that may be marvelous in itself but may also distort and disturb. One can imagine the discussions at the Bureau of Labor Statistics about these matters; I surmise that they are protracted and heated, yet they give rise to extraordinarily sober, almost inhuman, prose. Presumably such prose is used to help maintain objectivity and professional integrity, but one may wonder whether objectivity in such matters is ever possible. The very choice of things to mention, the order in which they are mentioned, the choice even of sober adjectives and adverbs, and so on, all have inevitable subjective elements. I suggest that a public document explaining how the Bureau of Labor Statistics (or any agency) goes about making the transition from numbers to words would be a valuable step toward that desirable, if never attainable, objectivity.

If summaries in words of tabulated statis-

tics could somehow be done by an algorithm, then we might examine the algorithm and make judgments about it. Alas, or perhaps hooray, no such algorithm seems likely. Variations in the background of the data and in the goals of the analysis stand in the way. Besides, how can an agency recruit and keep good analysts without giving them personal responsibility and opportunities for creative work?

In one approach by the National Center for Health Statistics, the text of a report is broken into separate statistical statements that are classified into five types. The first type is a simple characterization, for example: "The average length of illness for males was eight days"; the second type is a single comparison, for example: "Males were ill on average for two days longer than females." Then rules are applied related to underlying sampling or measurement error. In a simple characterization, for example, the coefficient of variation must be no more than 25 percent. For a single comparison, there must be statistical significance at the 5 percent level.

One reservation about such a procedure is that it atomizes a report into little single statements without much attention to pattern or interrelationships. One may also wonder about the effect of any set of rigid rules. Specification of a particular significance level, for example, seems constraining, even though editors of some scientific journals do it, and even though significance level is specified in one set of HEW equal economic opportunity regulations! A different kind of concern is about statements that *might* have been made but were not.

Statisticians know that words are important to statistics, yet surely their importance is not fully recognized. Self-conscious empirical studies of the meanings conveyed by variant ways of discussing statistics ought to be encouraged. It seems ludicrous for us to be concerned about exact significance levels, accuracy of standard errors, robustness of Bayesian procedures against fluctuation in prior beliefs, and so on, and yet to throw away all that exactness through our lack of concern about what gets communicated in the words we use—to each other and to our wider publics. Why do statisticians so seldom make studies of conveyed meanings? I do not know, but I take as a possible rallying cry a final quotation from Ralph Waldo Emerson, who says in his essay "The Poet": "Words and deeds are quite indifferent modes of the divine energy. Words are also actions, and actions are a kind of words."