

Extracting Valid Sexological Findings from Severely Flawed and Biased Population Samples

EDWARD M. BRECHER with JEREMY BRECHER

Since publication of the first Kinsey report (Kinsey, Pomeroy, & Martin, 1948), statisticians have repeatedly criticized human sexuality surveys on the ground that they are not based on probability samples. But sex surveys are not public opinion polls; they are inquiries utilizing the same scientific techniques relied on in the biological sciences generally: the initial sample of convenience, the comparison-group study, cumulative confirmations and deliberate sampling for heterogeneity, dose-response studies, the exclusion of confounding variables, the recognition of coherent patterns in the data, and others. These are powerful methodological instruments which do not require probability sampling for their efficacy. Probability sampling is needed primarily to answer questions beginning "What proportion of . . . ?" or "How many . . . ?" For answering such questions, nonprobability samples such as those used in sex surveys are not trustworthy; proportional information should therefore be eliminated from sex survey reports by means of indexing (normalization), except under special circumstances discussed here.

A statistical ghost has been haunting sex research ever since 1948, when Kinsey and his associates published their first report, *Sexual Behavior in the Human Male* (Kinsey et al., 1948). The purpose of this paper is to exorcise that ghost—to free sex research from a statistical straitjacket that has been hampering its progress for decades.

Shortly after the first Kinsey report appeared, statisticians rallied to attack it. A select committee of three was appointed by the American Statistical Association (ASA) to marshal these statistical criticisms, and the ASA thereafter published a 338-page critique entitled *Statistical Problems of the Kinsey Report* (Cochran, Mosteller, & Tukey, 1954). The committee members were much impressed with the Kinsey group's industry and thoroughness. They even paid the Kinsey report a backhanded compliment, calling it markedly superior to prior sex

Edward M. Brecher is the author of numerous publications in the field of human sexuality, including *Love, Sex, and Aging: A Consumers Union Report*. Jeremy Brecher, PhD, is an historian and author.

Requests for reprints should be sent to Edward M. Brecher, Yelping Hill, West Cornwall, CT 06796.

surveys. But on one point they were adamant: Kinsey's sample, they complained, was not a *probability sample*—that is, it was not drawn in such a way as to ensure that every American had an equal probability of being included in the sample. Instead, participants were, for the most part, self-selected volunteers and not representative of the U.S. population as a whole. Kinsey's findings must therefore be discounted as "subject to systematic errors of unknown magnitude due to selective sampling" (p. 37).

That same criticism has been exhumed and refurbished repeatedly as each new sex survey has made its appearance. The statisticians advancing it have academic credentials so impeccable and speak with such self-assurance that even some sex researchers have been brainwashed into believing that human sexuality surveys *should* be based on probability samples. Some sex survey reports even apologize for failing to meet the high statistical standards of a Gallup poll, a Roper poll, or a Harris poll!

I became personally involved in this sampling brouhaha back in 1977, when I was invited to direct a Consumers Union survey of geriatric sexuality from which ultimately emerged my book, *Love, Sex, and Aging: A Consumers Union Report* (Brecher & The Editors of *Consumer Reports Books*, 1984). I reviewed the relevant literature and was appalled by what I found. Those statisticians, despite their impeccable credentials, were clearly wrong. Probability samples are very useful for answering questions like "What proportion of voters are planning to vote for George Bush in 1988?" For answering many important kinds of scientific questions, however, a probability sample is not only difficult or impossible to secure but also unnecessary in principle and grossly inefficient in practice. There are many other routes to truth—methods in use throughout the sciences which do not depend on probability sampling.

Readers of this journal are no doubt already familiar with the elementary research methods reviewed below, such as comparison-group studies and the testing of hypotheses in widely disparate samples. But the way in which these universally accepted methods buttress the credibility of sex survey findings has almost wholly escaped notice;¹ hence this paper.

¹The National Institute of Mental Health, for example, sponsored a 1977 "Conference on Methodology in Sex Research"; the proceedings fill 332 pages (Green & Wiener, 1977). Except for a single sentence by Evelyn Hooker on page 297, none of the 27 well qualified participants in the conference, so far as the published record reveals, devoted any attention whatever to the methods reviewed in this paper. Hooker, after noting

The Sample of Convenience

Kinsey began his survey with a sample of which the only virtue was convenience: students in his Indiana University classes. Other sex researchers have followed his lead ever since—but with a guilty conscience. In fact, the sample of convenience is an appropriate starting point for many (perhaps most) studies in the biological sciences.

An example comes from an interdisciplinary research team that happened upon a parvovirus, subsequently dubbed RA-1, in the synovial tissue of a patient with rheumatoid arthritis (Simpson et al., 1984). They wondered, naturally, whether RA-1 was a chance contaminant in this particular synovial tissue specimen. So they examined 13 additional specimens that happened to be conveniently at hand—specimens from other rheumatoid arthritis patients who had just undergone joint surgery. Of the 14 specimens tested, 13 were positive for parvovirus RA-1.

A finding like this, based on a sample of convenience, hardly proves that RA-1 is a factor in the etiology of rheumatoid arthritis, but it may encourage a research team to seek out additional data. Kinsey's findings from his initial sample of convenience, indeed, encouraged him to devote the rest of his life to collecting additional data for the project he had so modestly launched among his Indiana University students.

The Comparison Group

Returning to the parvovirus RA-1 example, the next step was to examine synovial tissue specimens from the joints of a comparison group—in this case, eight patients with osteoarthritis instead of rheumatoid arthritis. No evidence of RA-1 was found in any of the comparison-group specimens. The score thus stood 13/14 for the presence of RA-1 in rheumatoid arthritis patients versus 0/8 in osteoarthritis patients. It did not take a statistician to assure this research team that the trail they were following might be leading toward a highly rewarding scientific pay-off, despite the fact that their sample so far contained only 22 patients and bore not the slightest resemblance to a probability sample.

To determine a linkage between two factors, the groups compared should be as alike as possible except for those two factors, such as type of arthritis and the presence or absence of RA-1 in the example above.

that it is "manifestly impossible" to use random sampling for a study of homosexuals, and "simply impossible" to use representative sampling, goes on to add: "Part of the solution is to study as many varieties of samples as possible." *Brava!*

If the two groups are both drawn from the same medical center or college campus, or even from the same classroom, this is a merit rather than a defect of the study protocol. If the two groups compared are drawn from a pool of self-selected volunteers, as in most human sexuality surveys, this does no harm—provided both groups selected themselves and volunteered under the same conditions.

Cumulative Findings

Once a potentially interesting finding has been suggested by a sample of convenience and has survived a comparison-group study, the appropriate next step is to determine whether the same finding emerges among groups as diverse as possible, an approach that has been labeled “deliberate sampling for heterogeneity” (Cook & Campbell, 1979, p. 75). In the parvovirus RA-1 study, the first heterogeneous sample came from England. There too it turned out that synovial tissue specimens from some rheumatoid arthritis patients, but not from osteoarthritis patients, showed evidence of RA-1 (Brown, 1984). This replication hardly solved the mystery of what causes rheumatoid arthritis, but the findings at this point were already sufficiently impressive to warrant publication in *Science* (Simpson et al., 1984) and *Nature* (Brown, 1984). No apology was made for the lack of a probability sample.

What distinguishes this RA-1 example from a Gallup or Harris poll is the kind of knowledge sought. The public opinion poll seeks primarily to answer questions beginning “What proportion of . . . ?” or “How many . . . ?” For this purpose, human sexuality surveys are of only limited value (see below). The RA-1 researchers, in contrast, were seeking to establish the existence of a linkage—a real-world linkage—between two previously unassociated factors, parvovirus RA-1 and rheumatoid arthritis. Many studies, including human sexuality studies, are of this general kind.

The Inherent Power of the Human Sexuality Survey

Since the first attacks on Kinsey, human sexuality surveys have been mistakenly likened to public opinion polls. Instead, as an example from Kinsey will demonstrate, the primary value of such surveys arises out of their character as linkage studies.

Kinsey discovered a linkage between delayed male puberty and low sexual frequencies in later years (Kinsey et al., 1948, p. 208). The shortcomings of his sample did not prevent him from making this

discovery, nor did it discredit his evidence for the existence of the linkage. To confirm this Kinsey finding, moreover, a probability sample would be much less efficient than a sample designed to maximize heterogeneity—one in which the presence (or absence) of the linkage could be demonstrated among husbands, widowers, divorced men, and bachelors; straight and gay; young, middle-aged, and old; and so on. A probability sample would be needed only to determine what *proportion* of males experience delayed puberty, or low sexual frequencies in later life, or for other proportional findings.

Table 1 presents selected data from my own survey, showing (among other things) a linkage between age by decade and the likelihood of remaining sexually active. Row A of Table 1 can be viewed as a report on a sample of convenience, composed of 1,145 wives aged 50 and over. Row B can then be viewed as a report on a comparison group composed of 1,895 husbands from the same sample pool. Both the Row A and the Row B groups are composed of *Consumer Reports* readers and their associates; both groups volunteered by filling out the same coupons and filling out and mailing back the same questionnaires. Thus, whatever their shortcomings as probability samples, the two groups are admirably qualified as comparison groups. The linkage

Table 1

Decade-by-Decade Changes in Proportion of Women and Men Who Are Sexually Active, and in Proportion of Sexually Active Women and Men Who Are Having Sex Once a Week or Oftener

	In their 50s	In their 60s	Aged 70 and over
	Proportion sexually active (masturbation or with a partner or both)		
A. Married women	(n = 590) 95%	(n = 495) 89%	(n = 160) 81%
B. Married men	(n = 643) 98%	(n = 811) 93%	(n = 441) 81%
C. Unmarried women	(n = 179) 88%	(n = 184) 63%	(n = 149) 50%
D. Unmarried men	(n = 152) 95%	(n = 131) 85%	(n = 131) 75%
	Proportion having sex once a week or oftener (masturbation or with a partner or both)		
E. Sexually active women	(n = 743) 73%	(n = 582) 63%	(n = 211) 50%
F. Sexually active men	(n = 804) 90%	(n = 893) 73%	(n = 473) 58%

Note. Adapted from Brecher et al. (1984), pp. 314-315.

indicated for wives in Row A is confirmed for husbands in Row B.

Rows C and D can next be viewed as cumulative evidence from a disparate sample, the *unmarried*. The same linkage between age and the likelihood of remaining sexually active is visible among the unmarried women and men in Rows C and D as among the wives and husbands in Rows A and B.

If there is in fact a real-world linkage between age and sexual activity, we would expect it to make its appearance with respect to *frequency* of sexual outlet as well as incidence. Rows E and F show that this is in fact the case for both sexually active women and sexually active men—thus providing further cumulative evidence for the linkage hypothesis.

The resources of a large-scale human sexuality survey do not end there, however. My data show a similar decline with age for the proportions of women and men who masturbate, for frequency of masturbation, for the proportions of wives and husbands who have sex with their spouses, for frequency of sex with spouses, and for level of enjoyment of sex among women and men (Brecher et al., 1984, pp. 315–316). If challenged, I could go on to show that the same linkage between age and incidence/frequency/enjoyment of sexual activity is visible among the rich and the poor, the pious and the nonreligious, high school dropouts and respondents with graduate degrees, in the East, Middle West, South, and Far West—all within the confines of my initial sample of 4,246 respondents. For this purpose, it is not necessary that each subgroup be represented in my sample in proportion to its presence in the U.S. population at large; it is only necessary that each subgroup compared be sufficiently numerous to warrant comparisons.

This explains, incidentally, why a major sex survey requires thousands of respondents, whereas Gallup or Harris dare to predict the outcome of a nationwide election with a sample much smaller than Kinsey's or mine. The larger sample is needed—Kinsey's goal was a sample of 100,000—in order to yield as heterogeneous a collection of subgroups and sub-subgroups as possible, each large enough to warrant comparisons and cumulative findings.²

²The conventional tests for statistical significance have been omitted from Table 1, as from Kinsey's statistical tables, for technical reasons falling outside the scope of this methodological primer. Instead of publishing tests of statistical significance, Kinsey and Brecher simply refrained from publishing data on groups too small to warrant comparisons. (For a further discussion, and especially for a detailed analysis of how large a group must be to warrant comparisons, see Kinsey et al., 1948, pp. 82–93 and Kinsey, Pomeroy, Martin, & Gebhard, 1953, p. 51.)

Reaching out to sample subgroups as heterogeneous as possible will often result in the over-representation of some subgroups, and therefore the under-representation of others. So be it. Sex researchers are not pollsters and should not try to be.

Indexing and Proportional Statements

Although there are some questions to which nonprobability samples like Kinsey's and mine cannot supply definitive answers, my Table 1 purports on its face to answer precisely such questions. The "95%" figure in the left-hand column of Row A, for example, purports to answer the question, "What proportion of wives in their 50s are sexually active?" Each of the 17 other percentages in Table 1 similarly purports to answer a "What proportion of . . . ?" question. In this respect, Table 1 is misleading—but the flaw can be readily remedied by a statistical technique known as *indexing* or *normalizing*, used in the biological sciences and demonstrated in Table 2.

Table 2 was prepared by assigning an index number, 1.00, to the group consisting of all of the respondents in Table 1. Each subgroup in

Table 2

Index Numbers^a for Decade-by-Decade Changes Shown in Table 1

	In their 50s	In their 60s	Aged 70 and over
	Sexually active (masturbation or with a partner or both)		
	(n = 590)	(n = 495)	(n = 160)
A. Married women	1.08	1.01	.92
	(n = 643)	(n = 811)	(n = 441)
B. Married men	1.11	1.06	.92
	(n = 179)	(n = 184)	(n = 149)
C. Unmarried women	1.00	.72	.57
	(n = 152)	(n = 131)	(n = 131)
D. Unmarried men	1.08	.97	.85
	Having sex once a week or oftener (masturbation or with a partner or both)		
	(n = 743)	(n = 582)	(n = 211)
E. Sexually active women	1.01	.88	.69
	(n = 804)	(n = 893)	(n = 473)
F. Sexually active men	1.25	1.01	.81

Note. Adapted from Table 1.

^aIndex number for upper portion of the table is based upon 4,066 participants; lower portion is based upon 3,706 participants.

Table 2 was then assigned an index number higher or lower than 1.00, depending on how that subgroup compares with the sample as a whole with respect to sexual activity and frequency. The percentage for each subgroup is divided by the percentage for the table as a whole to get the index number for that subgroup—in this case, $95\% \div 88\% = 1.08$. As a result, Table 2 presents precisely the same *pattern* as Table 1; but the suspect *percentages* have been eliminated by the indexing. The left-hand column of Row A in Table 2, for example, no longer alleges that 95% of wives in their 50s are sexually active; it now states only that the index number for this subgroup is 1.08, eight percent higher than the index number for the group as a whole—and similarly for all of the other data points in Table 2.

There are other methods of indexing, several of which—especially the “median split”³—can be very useful in sex research. If I were asked to identify the single most serious statistical shortcoming of human sexuality surveys from Kinsey (1948) to Brecher (1984), I would unhesitatingly cite the failure of sex researchers to exploit sufficiently the advantages of indexing their survey data.

There are some situations, however, in which a sex researcher may be warranted in publishing an answer to a “What proportion of . . . ?” question even though the answer is based on a nonprobability sample and is therefore less than trustworthy. An example of historic importance was Kinsey’s decision to publish his answer to the question, “What proportions of Americans are exclusively homosexual, exclusively heterosexual, or somewhere in between?” His decision had one justification: whatever the shortcomings of his sample, it was far and away the best sample ever assembled, or likely to be assembled in the then-foreseeable future, for answering that question. In the climate of his day, *some* estimates, even flawed estimates, were urgently needed. Kinsey’s estimates provided a far sounder basis for public policy than the ignorance and myths they superceded.

Perhaps a sound guideline is this: Sex researchers should in general eliminate proportional information from their statistical tables by means of indexing or in other ways. But when there is a genuine need for an answer to a specific proportional question, and when the data from a human sexuality survey are the best—or the only—data available, serious consideration should be given to publishing the answer. If

³For an explanation and example of the median split, see Brecher et al. (1984) pp. 262–263.

it is published, the shortcomings of the sample on which it is based should be explicitly noted.

Minimizing the Volunteer Effect

One reason that sex surveys cannot give definitive answers to "What proportion of . . . ?" questions is that the samples are composed of volunteers. How can we know whether those who fail or refuse to volunteer resemble those who did?

Three methods are available for measuring and perhaps discounting the volunteer effect:

1. Kinsey made very strenuous efforts to secure the participation of *all* of the members of some groups—all of the attenders at a church supper, for example. Among these 100% groups, there were no non-volunteers left out. He found that these 100% groups yielded substantially the same findings as the total sample (Kinsey et al., 1948, pp. 93-103).

2. Hunt invited persons selected at random from telephone directories to attend small-group discussions of recent trends in American sexual behavior; nothing was said about supplying personal sexual information. Thus, although those who attended the discussions were volunteers, they were not volunteers for a human sexuality survey. "Motivated by the discussion and unwilling to walk out on the group," Hunt reports, "virtually 100% of the discussants completed usable questionnaires" (Hunt, 1974, pp. 16-17).

3. Some surveys are based on mail questionnaires requesting participation. Those who fail to volunteer in this "first wave" then receive a second urgent plea for participation, citing additional reasons and perhaps offering rewards. A persuasive phone call may follow. By comparing the first wave with subsequent respondents, the extent of the volunteer effect can be roughly measured. Sometimes a weighted sample can be assembled, composed of appropriate numbers of first-wave and subsequent respondents.

Such techniques may be useful for placating critical statisticians, and they may improve the usefulness of a survey for answering "What proportion of . . . ?" questions. But they are very unlikely to affect scientific linkage findings:

4. If the linkage turned up in the study of a volunteer sample is in fact a real-world linkage, minimizing the volunteer effect will merely yield the same result over again.

5. If the hypothesized linkage does not exist in the real world, it is

very likely to be unmasked as spurious within the volunteer sample itself, by one or more of the methods reviewed above.

6. Differences between the initial all-volunteer sample and the subsequent sample are likely to be differences of degree rather than of kind; such differences are relevant to proportional questions but only rarely to linkage questions.

The Dose-Response Relationship and Experiments of Nature

Linkages visible in closely similar comparison groups and in cumulative findings among heterogeneous groups are further buttressed if a dose-response relationship can be discerned. A familiar example is the testing of a newly discovered chemical compound—call it Compound X—to determine its capacity to relieve pain. For this purpose, some patients in pain are given a *zero* dose of Compound X (that is, a placebo); others are given small, moderate, and large doses. If a unidirectional curve emerges, with the zero dose providing a baseline and each dosage increment yielding an increase in pain relief, the hypothesis that Compound X relieves pain is powerfully buttressed. In general, the evidence for linkage is buttressed if a little of Factor A yields a little of Factor B and a lot of A yields a lot of B. Note that some linkages are nonlinear; hence the absence of a unidirectional dose-response curve does not imply the absence of any linkage.

Tables 1 and 2 show how unidirectional dose-response curves can similarly buttress sex survey findings. These tables can be viewed as a report on an “experiment of nature,” in which nature rather than a human experimenter has varied the dosage among individuals in the sample. Thus, respondents in the column headed “In their 50s” received a minimum or baseline dose of aging, whereas “In their 60s” represents a dosage increment and “Aged 70 and Over” a further increment. The fact that the data points in Row A of the tables decline unidirectionally, step by step, from each age subgroup to the next buttresses the conclusion that age and sexual activity are in fact linked variables.

Dose-response studies, like comparison-group studies, can be repeatedly confirmed within a single sex-survey sample. Indeed, in Tables 1 and 2 the unidirectional downward curve for Row A is replicated by similar curves in Rows B, C, and D; and the curve of Row E is replicated by the curve of Row F. In fact, the difference between women and men with respect to sexual activity is *understated* in Rows A versus B and E versus F of Tables 1 and 2; for marital coitus is one

of the activities covered in the tables, and some of the women in Rows A and E are married to men in Rows B and F. The frequency of marital coitus is necessarily identical for a wife and her husband; this tends to dilute the sex differences shown in the table. This phenomenon, known as an interaction, can be a source of serious error. If the table covered *only* marital coitus between wives and their husbands, for example, it might lead unwary readers to conclude that there are no sex differences whatever with respect to either incidence or frequency of sexual activity. There are tests for determining the presence or absence of interactions but they fall outside the scope of this methodological primer.

Kinsey presented far more detailed and therefore even more persuasive dose-response curves. He divided his male sample into 18 age classes from 0-5 through 86-90, as well as into 10 occupational classes, 9 educational levels, 4 levels of religious commitment, and so on. He could thus show that a particular variable such as male frequency-of-orgasm declines step-by-step without exception through 14 successive age groups from 21-25 to 86-90. Moreover, he could confirm that unidirectional curve many times over, at various educational levels, among various occupational classes, and with various degrees of religious commitment. His achievement was indeed monumental.

The basic sampling rule for dose-response studies is the same as the rule for comparison-group studies: Participants at each dosage level should be selected, or should select themselves, in precisely the same way as those at other dosage levels.

The Exclusion of Confounding Variables

Even though the data in a sex survey appear to show a linkage between Factor A and Factor B, the two may in fact be independent. There may be a confounding variable, Factor X, which is linked to both A and B and may thus give rise to a spurious A-to-B linkage. In both the biological and the social sciences, confounding variables of this kind are a frequent source of serious error.

In Tables 1 and 2, for example, the apparent linkage between increasing age and declining sexual activity may in fact result from a confounding variable. As they grow older, people's health tends to worsen; perhaps it is really this increase in poor health that is responsible for the decline in sexual activity mistakenly attributed to aging *per se* in the tables.

Fortunately, there are highly effective ways to check for confound-

ing variables *within* a sex survey. In this case, I selected a "healthy comparison group" from my sample of 4,246—a group from which were excluded all respondents who rated their health as fair or poor or who failed to answer the question. I also excluded all respondents who reported removal of both ovaries, prostate trouble or prostate surgery, a heart attack, the taking of anti-hypertensive medication, and a number of other negative health factors increasingly common in the later years. Members of the healthy comparison group, as expected, showed a higher level of sexual function than those excluded from the group; but the unidirectional decade-by-decade decline in sexual activity with age shown in Tables 1 and 2—the dose-response curve—turned out to be similar for those in the healthy comparison group *and* for those excluded from the group. It held true for both healthy and health-impaired women and for both healthy and health-impaired men. Thus I could conclude with considerable assurance that the linkage between age and declining sexual function shown in Tables 1 and 2 is *not* a mere statistical artifact resulting from a confounding linkage between poor health and declining sexual function. If someone were to propose some factor other than impaired health as a potentially confounding variable, I could check that factor in the same way or with a more sophisticated statistical procedure such as multivariate analysis.

Note that using a probability sample does *not* protect against confounding variables, and that techniques which do protect are no more effective with a probability sample than with a nonrepresentative sample composed of self-selected volunteers. Here as elsewhere, the demand for probability sampling may be a red herring, distracting attention from methodological issues more relevant to the sex survey.

The Emergence of Coherent Patterns

Aspirin has been in use for many decades and has been the subject of hundreds of studies. Not a single one of those studies, taken by itself, establishes with a high degree of certainty that aspirin relieves pain. Nor can any one study demonstrate beyond cavil that parovirus RA-1 is important in the etiology of rheumatoid arthritis. But when all of the available studies are viewed together in perspective, a coherent pattern may emerge. Such a pattern provides the foundation on which important practical decisions, even life-or-death decisions, can prudently be based. The search for coherent patterns lies at the very heart of research in mathematics, in physics, in astronomy, in evolu-

tionary and ecological studies, and in other scientific endeavors—including human sexuality surveys.

Table 2 illustrates a coherent pattern emerging from my *Love, Sex, and Aging* data. The table exhibits the following six patterned regularities:

1. The index numbers for those in their 50s are higher than for those in their 60s in six out of six comparisons.
2. The index numbers for those in their 60s are higher than for those aged 70 and older in six out of six comparisons.
3. The index numbers for men are higher than those for women in eight out of nine comparisons.
4. The index numbers for the married are higher than those for the unmarried in six out of six comparisons.
5. The dose-response curves are unidirectionally descending in six out of six cases.
6. The data for sex frequency echo the data for incidence in six out of six cases.

If challenged, moreover, I could go on to show that the *same* coherent sixfold pattern emerges among the healthy and the health-impaired, the faithful and the adulterous, the urban and rural, and so on.

How can such patterned regularities be explained? If a large enough collection of monkeys is set to work striking typewriter keys at random for a long enough period of time, at least one of the monkeys will eventually and inevitably produce the full text of Shakespeare's *Macbeth*. No one, however, believes that *Macbeth* was ever in fact produced in that way—nor were the patterns emerging from the Kinsey and Brecher tables produced either by chance or by flaws and biases in our samples.

My Table 2 presents 18 data points arrayed in a remarkably coherent pattern. The table thus demonstrates on its face, as does *Macbeth*, that it was *not* produced by a monkey typing at random. Although the pattern displayed in those 18 orderly data points *might* have arisen by chance, or *might* have been produced by sampling vagaries which randomly raised some data points and lowered others—raising or lowering each data point precisely enough to produce this remarkably regular pattern despite the absence of such a pattern in the real world—those hypotheses can be dismissed as incredible for the same reason that we dismiss as incredible the hypothesis that *Macbeth* was typed by a monkey.

No, coherent patterns emerge from my data, and a fortiori from Kinsey's, for a very different reason. They are reflections in our data of real patterns and linkages present in the real world of real women and men. Indeed, the central purpose of the methods described in this paper is to distinguish real-world patterns from statistical artifacts as effectively as we can.

The Well Controlled, Double-Blind Prospective Experiment

The scientific methods here reviewed are not ideal. The ideal technique for establishing a linkage is the well controlled, double-blind, prospective experiment in which a sample is divided at random—for example, by flipping a coin—into an experimental group and a control group before the experiment begins. Then the experimental group is given the treatment under study, whereas the control group is given a placebo or else a treatment whose efficacy has already been determined. Neither the experimenters nor the subjects know, until the experiment is terminated, which group is receiving the treatment under study. If this protocol is strictly followed, whatever subsequently happens to members of the experimental group but not to members of the control group can with a very high degree of confidence be attributed to the treatment under study.

Human sexuality surveys are inferior to well controlled, double-blind, prospective experiments. They necessarily lack one essential feature: the *random* assignment of sample members to two groups *before* a treatment is administered. But note that the overwhelming bulk of modern science falls similarly short of this experimental ideal. The whole of astronomy, archaeology, meteorology, paleontology, and many other respected sciences lack even one example of a well controlled, double-blind, prospective experiment. Even in the medical sciences, the ancestral home of such experiments, only a trivial proportion of the science actually depended upon by physicians and surgeons is based on such experiments.

The fact that the human sexuality survey, like most of contemporary science, falls short of the methodological ideal hardly warrants a sexological inferiority complex.

The Statistical Straitjacket

Strange as it may seem, the impeccably qualified statisticians who drafted the ASA's critique of Kinsey's methodology paid little heed, anywhere in their 338-page report, to the scientific methods that

Kinsey actually used and I have here described. They neglected his comparison-group studies, his dose-response curves, his deliberate sampling for heterogeneity, his exclusion of confounding variables, and the emergence of meaningful patterns from his data.⁴ Instead, they measured Kinsey's shortcomings with the yardstick appropriate for public opinion polls—the probability sample yardstick. In the process, they blinded themselves to the powerful and unchallengeable scientific methods on which human sexuality surveys in fact rely.⁵ Other critics have viewed subsequent sex surveys from the same purblind perspective.

The founders of the science of statistics were innovative pioneers who sought to free science from the straitjacket of traditional logic—a logic obsessed in those days by if-then statements and by statements about all, some, or none of the members of a class. It is ironic that statistics today, obsessed by proportional statements requiring probability sampling, has itself become a straitjacket for sex research. I hope that this will prove less true hereafter.

⁴Kinsey's use of 100% samples was discussed within the committee, but the committee members were not placated.

⁵One member of the ASA committee commented informally that he would gladly trade Kinsey's 18,000 case histories for 100 gathered from a probability sample! (W. B. Pomeroy, personal communication, April 4, 1984)

References

- BRECHER, E. M., & the Editors of *Consumer Reports Books*. (1984). *Love, sex, and aging: A Consumers Union report*. Boston: Little, Brown.
- BROWN, K. A. (1984, June). Rheumatoid arthritis: Another viral candidate. *Nature* (London), p. 582.
- COCHRAN, W. G., MOSTELLER, F., & TUKEY, J. W., with the assistance of Jenkins, W. O. (1954). *Statistical problems of the Kinsey report on sexual behavior in the human male*. Washington, DC: American Statistical Association, Committee to Advise the National Research Council Committee for Research in Problems of Sex.
- COOK, T. D., & CAMPBELL, D. T. (1979). *Quasi-Experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.
- GREEN, R., & WIENER, J. (Eds.). (1977). *Methodology in sex research*. Bethesda, MD: National Institute of Mental Health.
- HUNT, M. (1974). *Sexual behavior in the 1970s*. New York: Dell.
- KINSEY, A. C., POMEROY, W. B., & MARTIN, C. E. (1948). *Sexual behavior in the human male*. Philadelphia: Saunders.
- KINSEY, A. C., POMEROY, W. B., MARTIN, C. E., & GEBHARD, P. H. (1953). *Sexual behavior in the human female*. Philadelphia: Saunders.
- SIMPSON, R. W., MCGINTY, L., SIMON, L., SMITH, C. A., GODZESKI, C. W., & BOYD, R. J. (March 30, 1984). Association of Parvoviruses with Rheumatoid Arthritis of Humans. *Science*, 1425-1428.