Randomization, Statistics, and Causal Inference

Sander Greenland

This paper reviews the role of statistics in causal inference. Special attention is given to the need for randomization to justify causal inferences from conventional statistics, and the need for random sampling to justify descriptive inferences. In most epidemiologic studies, randomization and random sampling play little or no role in the assembly of study cohorts. I therefore conclude that probabilistic interpretations of conventional statistics are rarely justified, and that such interpretations may encourage misinterpretation of nonrandomized studies. Possible remedies for this problem include deemphasizing inferential statistics in favor of data descriptors, and adopting statistical techniques based on more realistic probability models than those in common use. (Epidemiology 1990;1:421–429)

Keywords: statistics, causal inference, epidemiologic methods.

In this paper, I wish to review some ideas that, though of long history, seem too often overlooked when epidemiologists use statistical methods. My topic is the role of statistics in causal inference, but I am not here concerned with the issue of confidence intervals versus *P* values (an issue which, I think, was successfully resolved by Poole (1)). Nor am I here concerned with arguments favoring likelihood ratios over *P* values (2). Rather, my focus will be the meaning of these inferential statistics in randomized studies and their limited relevance when neither man nor nature has randomized the study exposure. The limitations, in particular, I feel are often underappreciated. My basic points are these: By continuing to define the computed values of inferential statistics in a probabilistic manner (as in virtually all texts), we encourage misinterpretation of the computed values when applied to nonrandomized studies. We also miss the opportunity to interpret them in a nonprobabilistic manner, and we overstate their importance for inference. We need to elevate the importance of data description and summarization relative to statistical inference, or else adopt inferential procedures based on more realistic probability models than current procedures; most likely, both remedies are warranted.

The arguments given here are largely derived from the writings of R.A. Fisher (eg, 3,4), Oscar Kempthorne (eg, 5), Jerome Cornfield (eg, 6,7), and David Freedman (eg, 8–10). Although these writers disagree (or would have disagreed) with each other on many points and I have not tried to reproduce faithfully their arguments or philosophies, each of them has attempted to clarify the meaning and limitations of inferential statistics when randomization assumptions fail to hold. Many other writers have put forth either parallel or dissenting views, but my intention is to review some logic rather than the literature (which is vast). I begin with a review of heuristic arguments leading to Fisher's exact test for two-by-two tables in randomized trials. Although all the salient points can be found in many textbooks, I wish to cover the arguments in some detail, in order to locate where the arguments break down for nonrandomized studies.

From the Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024-1772.

Randomized Trials

Randomization and Statistics

The statistical consequences of randomization may be illustrated as follows: Suppose I wish to study whether lidocaine prophylaxis prevents death within the 72 hours following hospital admission for acute myocardial infarction. I will enroll two patients for this study, two successive admissions to a hospital emergency room. When the first patient is admitted, I will toss a fair coin: If heads, the first patient will receive lidocaine and the second will not; if tails, the second admission will receive lidocaine and the first will not.

Suppose now that the first admission is massively compromised and is certain to die within 72 hours of admission, whereas the second is a mild case and is certain to survive, whether or not either of them receives lidocaine therapy. These conditions mean that lidocaine can have no effect on survival within this little cohort of two patients. In particular, even if both or neither of the patients were treated, we would observe exactly one half of the cohort die within 72 hours of

admission. Thus the "true" (causal) risk difference for lidocaine's effect in this cohort is ½ − ½ = 0. Given the preceding scenario, in which lidocaine has no effect, there are only two possible results of the randomized trial: If the coin is heads, the result will be

	Lidocaine		
	Yes	No	Total
Death			
Yes	1	0	1
No	0	1	1
	1	1	2

Risk-difference estimate = 1 − 0 = 1.

If the coin is tails, the result will be

	Lidocaine		
	Yes	No	Total
Death			
Yes	0	1	1
No	1	0	1
	1	1	2

Risk-difference estimate = 0 − 1 = −1.

(I call these "results" rather than "data" because they represent the actual experience of the subjects; "data" refers to observations that may not correctly represent subject experience, owing to losses, measurement error, etc.) Because of randomization, the probability of each of these results is ½. This fact has an interesting consequence for the risk-difference estimate: its mean (expected value) over the two possible results is

$$(½)(1) + (½)(-1) = 0,$$

exactly equal to the true risk difference. Thus we can see one statistical benefit of randomization: It makes our simple estimate of the true risk difference statistically unbiased, in that the statistical expectation (average) of the estimate over the possible results equals the true value.

Another benefit of the randomization is that it provides a known probability distribution for the possible results under a specified hypothesis about the treatment effect. From this probability distribution, we can calculate the standard error of the risk-difference estimate and an exact *P* value to "test" the specified hypothesis. Thus, in the above example, suppose we wish to test the null hypothesis (that the patients would have had the same outcome no matter how lidocaine was allocated). If we observe the first result, the lower *P* value is the probability of this result under the null hypothesis, which is ½, plus the probability of all possible results showing an association in a more negative direction; since there is only one more negative result (the second one) and its probability under the null is also ½, the lower *P* value is ½ + ½ = 1.

I chose the smallest possible controlled trial to illustrate these benefits, not only to keep the computations simple, but to illustrate one thing randomization does *not* do: It does *not* prevent the epidemiologic bias known as confounding. No matter what the outcome of randomization, the study will be completely confounded, in that the two treatment groups (comprising one patient each) will be completely noncomparable. In the first result, the greater severity of the first patient's condition is completely confounded with treatment, so that treatment appears highly causative of death (risk difference = 1); in the second result, the lesser severity of the second patient's condition is completely confounded with treatment, so that treatment appears highly preventive (risk difference = −1).

Confounding and Statistical Bias

The last example illustrates a basic discrepancy between the statistical concept of bias and the epidemiologic concept of confounding (when the latter is defined in terms of noncomparability or nonexchangeability of the compared groups (11)). Epidemiologic confounding is a property of an allocation, and for each allocation in the above example there is confounding in the extreme. In contrast, statistical bias refers to a nonzero *average* deviation over the probability distribution of results, and there is no statistical bias in the example.

Nevertheless, we can build a link between the two concepts if we measure the confounding in each result as the difference between the estimate and the true effect. For the first result, this measure of confounding is 1 − 0 = 1; for the second, it is −1 − 0 = −1. The average confounding across the results is thus (½)(1) + (½)(−1) = 0. This calculation illustrates a more general fact: The statistical unbiasedness of randomized trials corresponds to an average confounding of zero over the distribution of study results.

Randomization and Confounding

The preceding observation should provide little comfort for an epidemiologist trying to interpret a single result;

after all, what matters is the degree of confounding in the observed result. But randomization provides some indirect comfort if properly carried out: Using randomization, one can make the probability of severe confounding as small as one likes by increasing the size of the treatment cohorts. Of course, this fact is not an ironclad guarantee that confounding will not be severe, for our result could be one of the unlucky ones with severe confounding. Still, if our cohorts are large and we have no evidence of noncomparability, randomization should lead us to assign high credibility to the hypothesis that the cohorts are approximately comparable, provided there have been no gross violations of the assignment protocol (7).

To illustrate these points, suppose now that we enroll 1,000 consecutive admissions in our lidocaine trial, randomly allocating 500 of them to lidocaine prophylaxis, and suppose 100 of them die within 72 hours. Suppose further that allocation to lidocaine has no effect on the outcome of any of these patients, ie, the 100 that died would have died and the other 900 would have survived, regardless of allocation. Then, *regardless of allocation*, the study results will appear as follows:

	Lidocaine		
	Yes	No	Total
Death			
Yes	A	$100 - A$	100
No	$500 - A$	$400 + A$	900
	500	500	1,000

Note that under the hypothesized scenario (of no lidocaine effect), all the margins of this table are fixed.

In order to do the standard (frequentist) analysis of the table, we must calculate the proportion of allocations that lead to the observed table or a table showing a more extreme association. There are $\binom{1,000}{500} = \frac{1,000!}{500!(1,000 - 500)!}$ ways to allocate 500 of the 1,000 admissions to treatment. Of these allocations, there are $\binom{100}{A}$ that allocate A of the 100 deaths-to-be to treatment and $\binom{900}{500 - A}$ that allocate $500 - A$ of the 900 survivors to treatment, so there are $\binom{100}{A}\binom{900}{500 - A}$ allocations in which A deaths-to-be and $500 - A$ survivors receive the treatment. Thus, under the null scenario, the proportion of allocations in which A deaths-to-be and $500 - A$ survivors end up in the treatment group is

$$\frac{\binom{100}{A}\binom{900}{500 - A}}{\binom{1,000}{500}}. \quad (1)$$

So far, we have not invoked the assumption that allocation was random. This assumption implies that all the allocations are equally likely; this implication in turn implies that Eq 1 is not just a proportion, but is also the *probability* (under the null scenario) that A deaths-to-be and $500 - A$ survivors receive treatment. Using Eq 1, we can calculate the probability under the null scenario of any possible result of our large trial. In fact, Eq 1 is simply the hypergeometric probability that would be used in Fisher's exact test.

Under the null hypothesis, we should expect about half of the 100 deaths to occur among the treated patients. Suppose, however, that we observe 30 deaths among the treated patients. Given no treatment effect, the probability that random allocation would result in 30 or fewer of the 100 deaths occurring in the treatment group is

$$\sum_{i=0}^{30} \frac{\binom{100}{i}\binom{900}{500 - i}}{\binom{1,000}{500}} < 0.0001,$$

which is just the lower P value for Fisher's exact test. We can epidemiologically interpret this P value as the probability, under the null hypothesis, that randomization would yield a result with at least as much downward confounding as the observed result.

As amply discussed in the statistics literature, the utility of the preceding interpretations for epidemiologic inference are limited by the fact that they refer to results that might have been observed but were not, as well as what was observed (one can avoid this problem if one uses pure likelihood methods; see, for example, Ref 2). Nevertheless, if we randomize, we can find epidemiologic interpretations for certain classical inferential statistics, such as Fisher's P value and Pearson's chi-squared, as well as for various extensions of these statistics (12). In particular, we can give these statistics meaning in terms of classical epidemiologic biases, such as confounding. Can we do the same if we do not randomize? There is one special case in which the answer is

yes: If nature or circumstance resulted in what is essentially random allocation and we *knew* this was so, we could employ all the above interpretations of our statistics. But such "natural experiments" are rare. If, as usual, circumstance has not been so kind as to randomize, the answer is less encouraging.

Nonrandomized Studies

Statistics and Causal Hypotheses

Suppose as before that we wish to study lidocaine prophylaxis, and we examine two successive myocardial infarction admissions, but that our study is nonexperimental (ie, the attending physician allocates lidocaine treatment as he or she sees fit). If the result is

	Lidocaine		
	Yes	No	Total
Death			
Yes	1	0	1
No	0	1	1
	1	1	2

Risk-difference estimate = 1,

what are the implications of the null scenario (that lidocaine had no effect on the outcome of either of these patients)?

One implication is that, if the treatment status of these patients had been interchanged, we would have observed a complete reversal of the association. In other words, under the null scenario, our result is extremely sensitive to single interchanges of treatment status. In particular, our result is extremely sensitive to the attending physician's judgment regarding the needs of these patients.

Unfortunately, without randomization, the null scenario would usually not imply anything about the probability of any particular result. Only additional assumptions about physician behavior will allow one to make even qualitative statements about probabilities. Suppose, for example, that the attending physician employs lidocaine in precisely those cases in which death seems likely to ensue without it, and withholds it when survival seems assured regardless. If the physician is good at predicting outcomes, the above result would be much more probable than the result in which the survivor-to-be received lidocaine and the death-to-be did not. In the extreme, the above result is inevitable if the physician always administers lidocaine to deaths-to-be and never does so to survivors-to-be.

Such examples show that, in most nonrandomized studies, inferential statistics do not provide valid probability statements about treatment effects. *P* values, confidence limits, and likelihood ratios for causal parameters are calculated using the assumption that all interchanges of treatment status are equally probable outcomes of the processes determining treatment status (at least within strata of controlled factors). This assumption is warranted by proper randomization, but is rarely justifiable without it. How, then, are we to interpret statistics from nonrandomized studies?

When no randomization assumption is justified or even plausible, Meier has proposed (13) that conventional statistics should be interpreted as a "best-case" scenario with respect to apparent precision. Such an approach is helpful for cautious interpretation of imprecise studies. For example, if a nonrandomized study yields 95% relative-risk confidence limits of 0.5 and 5.0, we might say that, even if this study had been a randomized trial, there would still be considerable uncertainty about treatment effect. Unfortunately, this approach is not of much use if the results appear sharp. For example, if the 95% limits for the relative risk were 4.0 and 6.0, this approach would only emphasize how little uncertainty would have remained if this study had been a randomized trial.

Statistics and Descriptive Hypotheses

Another approach to inferential statistics in nonrandomized studies is to regard them as "descriptive," in the following sense: One imagines that the treated and untreated groups represent random samples from two separate treated and untreated parent populations; the statistics refer to the difference in outcome frequency in the two parent populations but have no connection to any causal interpretation of this difference. For example, under the descriptive interpretation, the *P* value refers to the null hypothesis that the outcome frequencies in the treated and untreated parent populations are the same. The *P* value may be very small and this null hypothesis rejected, with no implication that the treatment has an effect. This interpretation allows for the possibility that the outcome frequencies in the populations may be different because of differences in the processes that generated the two parent populations. In other words, a comparison of the two parent populations might be confounded by exactly the same mechanisms that would confound comparisons within the study, such as selective allocation of high-risk patients to treatment. In

such circumstances, the statistics only aid us in inferring the structure of the parent populations, as in sample surveys.

Insofar as the descriptive interpretation restrains causal inferences from nonrandomized studies, it is a good thing. Nevertheless, the descriptive interpretation is rarely justified, for the simple reason that study (sample) cohorts are rarely based on random samples from any parent population. Consider the Framingham (Massachusetts) study of heart disease. Even if the investigators had achieved full participation and follow-up of the selected subjects (which they did not (14)), to what parent population would their statistics refer? Suppose one answers that, say, white Framingham males born in 1900 were a random sample of all white American males born in 1900 and alive at study inception (1948), so the statistics refer to the latter population. This claim must ignore the large ethnic heterogeneity across regions of the United States, such as the frequent Anglo-Irish background found in Framingham men, compared with the frequent Scandinavian backgrounds for Minneapolis men or the frequent German and Polish backgrounds for Milwaukee men. In light of the large variation in heart-disease rates across European nationalities, it would seem cavalier to ignore these differences and regard the Framingham cohort as a random sample from the general U.S. population.

One might then attempt to salvage the descriptive interpretation by restricting the definition of the parent population. One might restrict geographically, say, by claiming the statistics for white males born in 1900 refer to eastern Massachusetts men. But this claim must ignore the patchwork variation that existed in 1948 within eastern Massachusetts: Neither the English-descent "bluebloods" from the North Shore nor the relatively new Greek immigrants within Boston stood much chance of representation in the Framingham cohort; how could the Framingham cohort be a random sample of the population that included those ethnic groups?

Suppose we continued this line of reasoning, restricting the parent population to reasonably similar communities, and maintaining geographic continuity of the areas. We would soon find that any reasonable parent population for the Framingham cohort would have quite an artificial appearance and might not be very much larger than the town of Framingham! The parent population would also have to be quite restricted in time, given the profound and incompletely explained secular trends in heart disease.

Consider the very low *P* value observed in Framingham for the smoking–myocardial infarction association. Since smoking was not randomized, this *P* value could not refer to the causal null hypothesis that "smoking is not a cause of myocardial infarctions." Descriptively, one might claim the *P* value referred to the null hypothesis that "myocardial infarction rates are constant across smoking levels"; but the parent population to which this hypothesis refers would at best be some aggregation of "Framinghamlike" communities, with areas gerrymandered so that the random-sampling assumption would not be obviously false. It is not clear that imprecise information about this artificial and ill-defined population, such as relative-risk confidence intervals for the population, should be of more interest than precise information about the actual study cohort, such as the observed relative risks for the cohort.

If we now consider nonresponse and loss to follow-up in Framingham (which amounted to roughly one third of those originally invited to participate), the descriptive interpretation of the inferential statistics breaks down completely. The subjects did not make their decisions to participate or not on the basis of some random-sampling device. Consequently, the Framingham cohort experience that was actually observed is not a random sample of any parent population experience, such as a gerrymandered collection of "Framinghamlike" communities; it is not even a sample of the persons initially selected for study. One could attempt to salvage the descriptive interpretation by arguing that nonresponse and loss to follow-up were unrelated to smoking and heart disease, but this proposition (which is doubtful) is by definition unverifiable. After all, if we could observe that nonresponders and dropouts had the same joint distribution of smoking and heart disease as the retained part of the cohort, the nonresponders and dropouts would no longer be lost (15). Thus, in the Framingham study, the descriptive (sample survey) interpretation of the smoking–heart-disease *P* value turns out to be as groundless as the causal interpretation.

The point of the preceding exercise is not to criticize the Framingham study; on the contrary, it is important to note that the study was among the most informative in epidemiologic history. The point is that the study was informative despite the fact that the study statistics bore no randomization interpretation (since no one was randomized), and that any defensible descriptive interpretation would have to be trivial in character.

Statistics and Stochastic Models

Neither the causal nor the descriptive interpretations of inferential statistics holds up in typical nonrandomized

studies. But the descriptive interpretation has given birth to a third approach, which I would call "stochastic modeling." The general idea is this: We regard each individual's outcome as a random variable whose distribution depends only on treatment and (possibly) measured covariates; we then assume that this dependence has some simple form. Returning to the lidocaine example, we might regard each patient's survival as partly a matter of "chance," in exactly the same sense that the outcome of the coin toss is a matter of "chance." In other words, each patient is assumed to have his or her own survival probability, which may fall between zero and one.

Suppose we assume that the survival probabilities among treated patients in our study all equal a common value π_1, and the probabilities among untreated patients equal a common value π_0. Using these homogeneity assumptions, along with some criteria for choosing "best" tests, it is possible to deduce the lower Fisher P value as a test statistic for the one-sided null hypothesis $\pi_1 \leq \pi_0$ (that treated patients have no better chance of survival than untreated patients) (16). More generally, one can derive all the usual 2×2 table statistics for noncausal comparisons of π_1 and π_0, subject to the assumption of homogeneous probabilities within treatment groups.

A moment's reflection should reveal that the homogeneity assumption is absurd. If there is even one prognostic factor (eg, age) that varies within treatment groups, the survival probabilities will vary within groups and the assumption will fail. One could attempt to get around the assumption by stratifying on measured prognostic factors to create homogeneous subgroups; nevertheless, the usual stratified statistics would still depend on the homogeneity assumption holding within each stratum. Unfortunately, in most if not all settings, this stratified homogeneity assumption would be difficult to justify, since there are few if any epidemiologic settings in which all strong risk factors (such as susceptibility genes) are accurately measured and controlled. I would conclude that a stochastic-model interpretation cannot justify conventional statistics in nonrandomized studies.

Statistics and Data Description

As a final attempt to justify conventional statistics in nonrandomized studies, one might consider whether they have any nonprobabilistic justification as data descriptions or data summaries. It turns out that such justification is possible (8,17). In particular, conventional statistics may help give us a sense of how much our results would be influenced if a small proportion of the allocations were changed, at least if the number of subjects is so large as to preclude subject-by-subject consideration of interchanges. To illustrate this point, suppose we enroll 1,000 consecutive admissions into our nonrandomized lidocaine study, and observe

	Lidocaine		
	Yes	No	Total
Death			
Yes	30	70	100
No	470	430	900
	500	500	1,000

This result would hardly be influenced by making a few interchanges, and thus would be insensitive to a few changes in the decisions of the attending physicians. We may additionally note that, under the null scenario (that the outcomes of these patients were unaffected by treatment), the same combinatorial arguments as used in the randomized-study example may be used to show that fewer than 0.0001 of the $\binom{1,000}{500}$ possible interchanges of treatment status would lead to as much or more of an association in the preventive direction. Put another way, if the null scenario is correct, fewer than 0.0001 of the possible allocations could result in such an extreme or more extreme negative association (8).

Although the nonprobabilistic interpretation of classical statistics may form a basis for their use in nonrandomized studies of causation, several cautions apply to such use. In the preceding example, the P value told us that the allocation was extreme, but still did not tell us the *probability* of getting such an extreme allocation under the null. Unlike the randomized trial (in which all allocations are equally probable), this probability depends on a host of factors such as physician behavior. If, for example, the attending physicians preferentially apply lidocaine therapy to cases with good prognosis, a particular allocation that results in a negative association will have a higher probability than a particular allocation that results in a zero or positive association. This scenario implies that the true probability of observing a result as extreme as or more extreme than above is higher, perhaps much higher, than the value obtained from the classical calculation.

Even when classical statistics are considered as only data descriptions, it is easy to misinterpret them, especially if one of the treatment or outcomes groups is small. Consider for example, a nonrandomized study of lidocaine, with ten consecutive admissions and the following result:

	Lidocaine	
	Yes	No
Death		
Yes	0	5
No	5	0

The lower Fisher's exact P value is $1/252 = 0.004$ for this table, meaning that under the null scenario only 0.4% of allocations (in fact, only this allocation) would produce at least as extreme a negative association. Although 0.004 is a very small proportion, it does *not* mean the results are insensitive to a few changes in allocation (18). Interchanging one treated with one untreated patient represents a change in treatment status for 20% of the subjects. Under the null hypothesis, a single interchange will yield

	Lidocaine	
	Yes	No
Death		
Yes	1	4
No	4	1

Although the association is still strong, the exact P value is now 0.10. One more interchange of status could almost obliterate the association. Thus, if a single interchange involves a large proportion of a treatment or an outcome group, a very small (or very large) P value does not mean that a finding is insensitive to interchanges.

Alternatives to Classical Statistics

I have argued that, in most epidemiologic studies, one cannot justify classical statistics by any of the common probabilistic arguments. Given this predicament, what alternatives to classical statistics are available? One alternative is to limit statistical analysis to data description, at least when no agreed-upon probability model is available. Many innovative techniques for data description have appeared under the rubric of "exploratory data analysis" (19), especially techniques oriented toward creating informative visual summaries such as graphs and charts. Such summaries may be descriptive of the data only, as in basic tables or graphs of rates. Other tabular and graphical summaries, such as scatterplot smoothers (20), may be derived under random-sampling assumptions but can also be viewed as data descriptions.

An extension of data description is *influence analysis*, in which one explores the degree to which one's data summaries or effect estimates would change under small perturbations of the data (21,22). More generally, the most telling aspect of the stability of a result (especially in small studies) may be the impact of deleting or interchanging a few subjects, as in the last example of the last section. Influence analysis may of course be applied to randomized studies, but the fact that it can be used even when one has no probability model makes it especially suitable for nonrandomized studies.

Rather than discard probabilistic statistics when one knows of no random mechanism generating the data, one can take the opposite course of employing more elaborate probabilistic models. Starting from the basic models appropriate for randomized studies, one can add model terms for the effects of measured covariates (such as regression coefficients), unknown covariates (such as random effects), and for known sources of bias (such as correction terms for unmeasured confounders, selection bias, or measurement error).

The problems with modeling have been discussed at length in the literature, with the fundamental criticism being the dependence of modeling results on the correctness of the assumed model (9,10,23–25). This problem nevertheless arises in interpretation of any inferential statistics, including such elementary statistics as Mantel-Haenszel tests and confidence limits, as well as Fisher's exact test. More elaborate models have the virtue of explicitly accounting for known deviations from the ideal randomized study (the ideal under which elementary statistics are derived). Their drawback is of course their reliance on very detailed assumptions about processes (such as covariate effects) when there is little basis for such assumptions.

One response to the last problem is to conduct a *sensitivity analysis*, in which the analysis models (and hence the assumptions) are systematically varied to identify those findings (if any) that are relatively unaffected by model choice (26,27). Many investigators already employ an informal sensitivity analysis, insofar as they apply a variety of analytic techniques to their data to identify findings that emerge under every technique. (This "serial" method of evaluating findings should be contrasted to the potentially biased "parallel" approach, in which a finding is considered "real" if it emerges from just one of many techniques.) With formalized sensitivity analysis, one can better ensure exploration of a broad range of models.

Unfortunately, the range of plausible models will often be far too broad for full exploration. One theoretical answer to this problem is to adopt a Bayesian viewpoint: We may regard all models as points in a very high-dimensional parameter space, specify prior distributions

over that space, and see what inferences follow for any reasonable prior (26,27). I find this viewpoint invaluable for criticism of models and model-selection procedures (eg, see Ref 23), but it does not seem to lead to easily implemented or generally acceptable methods for data analysis.

Another proposed answer to uncertainty about model specification is to employ robust procedures (that is, procedures that work better than standard procedures when the assumptions underlying standard procedures are violated). Several approaches show promise in certain epidemiologic applications. Random-effect (frailty) models have been employed when the usual probability models for disease occurrence (such as Poisson or binomial) appear suspect (28–30); empirical Bayes methods may be employed to deal with uncertainties about structural models (31), and have already been used in several important epidemiologic problems such as disease mapping (32), smoothing of unstable rates (33), and screening of multiple associations (34,35). Nevertheless, one should bear in mind that no procedure is robust to all conceivable violations of the underlying probability model; in particular, any procedure will be biased by uncontrolled sources of confounding, differential selection, and measurement error.

Conclusion

Randomization provides the key link between inferential statistics and causal parameters. Inferential statistics, such as *P* values, confidence intervals, and likelihood ratios, have very limited meaning in causal analysis when the mechanism of exposure assignment is largely unknown or is known to be nonrandom. It is my impression that such statistics are often given a weight of authority appropriate only in randomized studies. As an example of this improper weight, my co-authors and I have often presented confidence intervals with a comment that "the data would appear to be compatible with a relative risk ranging from [lower limit] to [upper limit]." Careful consideration of the points raised here might have led us to at least preface such a comment with the phrase "if the exposure had been randomized within levels of the controlled variables, . . . " or else omit commenting on the inferential statistics altogether, since the data were probably compatible with a broader range of values than indicated by the limits. Parallel criticisms apply to the use of statistics for making descriptive inferences from samples to populations when the selection mechanism is unknown or known to be nonrandom, because random sampling provides the key link between inferential statistics and population parameters.

While a good argument can be made for the value of statistics in restraining our interpretation of data (36), many of us have come to rely on a limited body of statistical techniques to quantify the compatibility or conflict between data and a hypothesized effect, and the relative support of the data for different hypotheses. In causal analysis of observational data, valid use of inferential statistics as measures of compatibility, conflict, or support depends crucially on randomization assumptions about which we are at best agnostic and more usually doubtful. Among the possible remedies are: (a) Restrain our interpretation of classical statistics by explicating and criticizing any randomization assumptions that are necessary for probabilistic interpretations; (b) train our students and retrain ourselves to focus on nonprobabilistic interpretations of inferential statistics; (c) deemphasize inferential statistics in favor of pure data descriptors, such as graphs and tables; (d) expand our analytic repertoire to include more elaborate techniques that depend on assumptions in the "agnostic" rather than the "doubtful" realm, and subject the results of these techniques to influence and sensitivity analysis. These are neither mutually exclusive nor exhaustive possibilities, but I think any one of them would constitute an improvement over much of what we have done in the past.

Acknowledgments

I would like to thank Anders Ahlbom, David Freedman, Stephan Lanes, Charles Poole, and Jan Vandenbroucke for their helpful comments on the original manuscript. I am also indebted to James Robins, who has helped me immensely in thinking about these issues.

References

1. Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195–9.
2. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988;78:1568–74.
3. Fisher RA. The design of experiments. 6th ed. New York: Hafner, 1951.
4. Fisher RA. Statistical methods and scientific inference. 3rd ed. New York: Hafner, 1973.
5. Kempthorne O. Comment on "The Bayesian outlook and its applications" by J. Cornfield. Biometrics 1969;25:647–54.
6. Cornfield J. Principles of research. Am J Mental Deficiency 1959;64:240–52.
7. Cornfield J. Recent methodological contributions to clinical trials. Am J Epidemiol 1976;104:408–24.
8. Freedman DA, Lane D. Significance testing in a nonstochastic setting. In: Bickel P, Doksum K, Hodges JL, eds. Lehman Festschrift. Belmont, CA: Wadsworth, 1983.
9. Freedman DA. Statistics and the scientific method. In: Mason W, Feinberg SE, eds. Cohort analysis in social research: beyond the identification problem. New York: Springer, 1985:345–90.
10. Freedman DA. As others see us: a case study in path analysis (with discussion). J Educ Stat 1987;12:101–223.
11. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiologic confounding. Int J Epidemiol 1986;15:412–18.

12. Robins JM. Confidence intervals for causal parameters. Statist Med 1988;7:773–85.
13. Meier P. Damned liars and expert witnesses. J Am Statist Assoc 1986;81:269–76.
14. Gordon T, Moore FE, Shurtleff D, et al. Some methodologic problems in the long-term study of cardiovascular disease: observations on the Framingham study. J Chron Dis 1959;10:186–206.
15. Greenland S. Response and follow-up bias in cohort studies. Am J Epidemiol 1977;106:184–7.
16. Lehman EL. Testing statistical hypotheses. 2nd ed. New York: Wiley, 1986.
17. Rouanet H, Bernard J-M, LeCoutre B. Nonprobabilistic statistical inference: a set theoretic approach. Am Statist 1986;40:60–5.
18. Dupont WD. Sensitivity of Fisher's exact test to minor perturbations in 2 × 2 contingency tables. Statist Med 1986;5:629–35.
19. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977.
20. Hastie T, Tibshirani R. Generalized additive models. Statist Sci 1986;1:297–318.
21. Cook RD, Weisberg S. Residuals and influence in regression. New York: Chapman and Hall, 1982.
22. Pregibon D. Data analytic methods for matched case-control studies. Biometrics 1984;40:639–51.
23. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986;123: 392–402.
24. Vandenbroucke J. Should we abandon modeling altogether? Am J Epidemiol 1987;126:10–3.
25. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;340–9.
26. Leamer EE. Specification searches. New York: Wiley, 1978.
27. Leamer EE. Sensitivity analyses would help. Am Econ Rev 1985;75:308–13.
28. Williams DA. Extra-binomial variation in logistic linear models. Applied Statist 1982;31:144–8.
29. Breslow NE. Extra-Poisson variation in log-linear models. Applied Statist 1984;33:38–44.
30. Aalen OO. Heterogeneity in survival analysis. Statist Med 1988;7:1121–37.
31. Morris CN. Parametric empirical Bayes inference: theory and applications (with discussion). J Am Statist Assoc 1983;78:47–65.
32. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987;43:671–81.
33. Tsutukawa RK, Shoop GL, Marienfeld CJ. Empirical Bayes estimation of cancer mortality rates. Statist Med 1985;4:201–12.
34. Thomas DC, Siemiatycki J, Dewar R, et al. The problem of multiple inference in studies designed to generate hypotheses. Am J Epidemiol 1985;122:1080–95.
35. Greenland S. A semi-Bayes approach to the analysis of correlated multiple associations, with an application to an occupational cancer-mortality study. Technical Report No. 1, Department of Epidemiology, UCLA School of Public Health, 1990.
36. Diaconis P. Theories of data analysis: from magical thinking through classical statistics. In: Hoaglin DC, Mosteller F, Tukey JW, eds. Exploring data tables, trends, and shapes. New York: Wiley, 1985.