Book Reviews

Am. J. Hum. Genet. 63:283–289, 1998

Statistical Evidence: A Likelihood Paradigm. By Richard Royall. London: Chapman & Hall, 1997. Pp. 191. $64.95

What follows is a book review, in a somewhat unconventional format. The book under consideration is about the foundations of statistical inference, and it may not be clear to many human genetics researchers that they should care about such matters. We think that they should, and this essay is our attempt to explain why. For those who find our comments intriguing or for those with an existing interest in this area, Richard Royall's new book *Statistical Evidence: A Likelihood Paradigm* will be worthwhile reading.

Trouble in Paradise

Research into the genetic basis of human diseases takes place at the intersection of three distinct fields: medicine, molecular biology, and statistics. Although most human geneticists are conversant in all three areas, specialists from each discipline nevertheless depend, for complementary expertise, on their colleagues from the others. It may not come as a welcome surprise to many clinicians and molecular geneticists, therefore, to learn that all is not well in the house of statistics: there exist deep philosophical divisions within the field, and, perhaps surprisingly, this has direct bearing on the conduct of human genetic studies.

As an illustration, consider the recent debate over genomewide significance levels in the human genetics literature: Is it more appropriate to report *P* values adjusted for the number of tests that would be carried out in the course of a complete genomic screen (regardless of whether an entire screen is actually performed) or simply to indicate the pointwise significance level for any tests actually performed (which allows the interested reader to correct for the number of tests actually performed)? On the one hand, it is argued that, if investigators fail to obtain statistical significance early on in a study, they almost invariably will proceed to continue screening the entire genome until a finding is obtained. Then failure to adjust for genomewide significance levels unfairly penalizes those investigators whose findings happen, by chance alone, to come late in their studies (Lander and Kruglyak 1995, 1996). On the other hand, it is argued that the investigator who in fact conducts only a limited number of tests is unfairly penalized if she is then made to "correct" for arbitrarily many tests that she could have but did not perform (Witte et al. 1996; also see Thomson 1994; Curtis 1996). Both sides seem to have valid points, and clinical investigators may be disturbed by the appearance of a dilemma having no clear resolution.

But there is a school of statistical thought that makes it unnecessary to choose between these equally unsatisfactory options. According to this school, if what we are really interested in is gauging the evidence for (or against) linkage, then the *P* value is not the proper measure to begin with, and the debate over multiple-testing adjustments to the observed *P* value becomes irrelevant. For those who are troubled by the multiple-testing dilemma, this school of thought provides an appealing alternative. In fact, this school of thought raises serious questions about some practices that may *not* be troubling clinical investigators but that perhaps should be—for example, the use of "model-free" tests in linkage studies.

An Extravagant Claim

In *Statistical Evidence,* Royall, a professor of biostatistics at Johns Hopkins University, offers a cogent and compelling defense of this other school of thought. The "likelihood paradigm" of Royall's subtitle will be familiar to those geneticists already acquainted with the earlier work of Edwards (1972). Since no general term exists for proponents of this school, we take the liberty of dubbing them "statistical evidentialists," and we will call the methods that they advocate "evidentialism." By contrast, the two other prominent schools are often referred to as "frequentists," on the one hand, and as "Bayesians," on the other.

Most of current statistical practice is based on frequentist principles—notably, on the Neyman-Pearson paradigm for hypothesis testing or on Fisher's conception of significance testing. Evidentialism is undoubtedly the least familiar school of statistical thought, both within the field of statistics itself and, certainly, among consumers of the statistical literature. This remains true, at least in part, because journal editors and peer reviewers almost invariably ask that statistical results be reported in familiar frequentist terms. But our predilection for the familiar notwithstanding, *evidentialism is, arguably, the only body of statistical theory that is fully consistent with the practice of science.*

The Problem with P *Values*

To justify this extravagant claim, we need to consider the *purpose* of statistical analysis in scientific contexts. Evidentialism views the purpose of statistical inference as the *measurement of the strength of evidence* conferred by a given set of data in favor of one hypothesis over another. This may seem a wholly natural objective for scientific data analysis, and we will take it as given that this is the objective that we are pur-

suing. But, in fact, much of standard statistical practice is based on a quite different conception of statistical inference—namely, as a set of tools for *decision making* in the face of uncertainty. This latter objective need not in any way involve the concept of evidence.

For example, the familiar (Neyman-Pearson) paradigm taught in all introductory statistics courses proceeds, in broad strokes, as follows: We begin with a "null" hypothesis and select an acceptable level of significance, which is the probability that we will reject the null hypothesis when it is in fact true (the type I error rate). We then select the "best" testing procedure, one that minimizes the probability that we will fail to reject the null hypothesis when it is in fact false (the type II error rate) for the selected significance level. Then we perform the test, reporting as results the type I error rate, the type II error rate, and our *decision* either to reject the null hypothesis at the chosen significance level or not to reject it. This overview of statistical inference may be so familiar as to also seem wholly natural. But nowhere does this account mention the measurement of evidence. Are the two error rates of Neyman-Pearson tests related to statistical evidence in some way?

The type I and type II error rates of a Neyman-Pearson test represent the frequencies with which certain events (the two types of error) will occur over repeated applications of the decision-making procedure—that is, they reflect our *prediction* of how the test procedure will behave, in general, prior to the collection of any data. But, once we have the data in hand, we are really more interested in how the test procedure happened to have performed in this particular application, which is quite another matter. To draw a loose analogy: prior to leaving my house in the morning, I might listen to the weather forecast before deciding whether to carry my umbrella; but, once I have left the house and find myself in a downpour, the fact that rain was predicted to be only moderately likely does not mitigate my regret at having left the umbrella at home. Similarly, once we have data in hand, we are no longer satisfied with reporting the probability that a certain erroneous outcome might occur when we perform a test of this sort. Rather, we would like to have some way to determine whether we have been misled in this instance. The predetermined significance level of a Neyman-Pearson test does not give us this information.

For this reason, in the scientific literature we hardly ever conform to the strict Neyman-Pearson paradigm in the reporting of statistical results. Rather, what we often see reported is the "empirical" *P* value of the test—that is, the type I–error probability corresponding to the observed value of the test statistic—rather than the simple decision to reject (or not to reject) the null hypothesis made at a predetermined level of significance. This empirical *P* value is then commonly interpreted as if it were a measure of the strength of the evidence, with smaller *P* values interpreted as reflecting greater evidence against the null hypothesis. The practice of reporting empirical *P* values reflects the fact that our objective is the measurement of evidence, rather than decision making per se. But, at the same time, using the empirical *P* value in this way is an attempt to address this evidentialist objective within the familiar frequentist decision-theoretic (hypothesis-testing) framework. This practice reflects our interest in *evidence,* while restricting our statistical focus to the predictive *error rates* of hypothesis-testing procedures.

But can the *P* value be made to do double duty, both as the predictive type I–error probability and as a measure of the strength of the evidence? What is the relationship between the question of statistical evidence and the frequentist's interest in error rates? Are these really just two ways of naming the same statistical quantities, or are these fundamentally different kinds of quantities? And, if the *P* value is not the appropriate measure of the strength of evidence, then what is? Although these questions might seem too philosophical to require the attention of genetics researchers, the methods that we choose for analysis of genetic data ought perhaps to depend on the answers that we give. The evidentialist's answers begin with the recognition that the familiar frequentist methods cannot be made to satisfy our interest in the measurement of evidence.

What Else Is There?

If we are to forgo the familiar frequentist methods, what alternatives do we have? Evidentialism offers an alternative paradigm that is expressly designed to address the problem of how best to measure statistical evidence. The cornerstone of this alternative paradigm is a definition of statistical evidence that is based solely on the likelihood ratio (LR). Royall puts this in terms of Hacking's (1965) "law of likelihood": "If hypothesis *A* implies that the probability that a random variable *X* takes the value *x* is $p_A(x)$, while hypothesis *B* implies that the probability is $p_B(x)$, then the observation $X=x$ is evidence supporting *A* over *B* if and only if $p_A(x) > p_B(x)$, and the likelihood ratio, $p_A(x)/p_B(x)$, measures the strength of that evidence" (p. 3).

As Royall says, the law of likelihood defines statistical evidence in a way that seems both "objective and fair," since it says, in essence, that "the hypothesis that assigned the greater probability to the observation did a better job of predicting what actually happened, so it is better supported by that observation. If the likelihood ratio, $p_A(x)/p_B(x)$, is very large, then hypothesis *A* did a much better job than *B* of predicting which value *X* would take, and the observation $X = x$ is very strong evidence for *A* versus *B*" (p. 5). This outlook does indeed seem sensible. The surprising thing is just how easily this law may be violated when we attempt to use the *P* value as a measure of evidence.

To illustrate this point, suppose that we are interested in establishing whether a given coin is "fair." In particular, to keep things simple, suppose that we are interested in determining whether the coin is truly disposed to land heads with probability $p_1 = \frac{1}{2}$ or probability $p_2 = \frac{1}{4}$. One experiment that we might perform would be to toss the coin a fixed number of times, *N,* and to record the number of times, *H,* that the coin lands heads. In this experiment, as is well known, the random variable *H* has a binomial probability distribution. Another experiment that we might perform would be to toss the coin repeatedly until the first time that it lands heads, recording the number, *M,* of tosses required. In this case, the random variable *M* has a geometric probability distribution. Both experiments yield information about the true underlying probability *p* that the coin will land heads.

Note that in this example we are explicitly concerned with comparing two hypotheses, $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{4}$. Some statisticians might prefer to talk about testing a "null" hypothesis

without reference to an alternative hypothesis. As we have already seen, however, the law of likelihood expressly applies to comparisons between two hypotheses: evidence counts against one hypothesis only insofar as it favors the other. This insistence that any proper measure of evidence must involve two hypotheses rather than one is a cornerstone of evidentialist theory. In the interest of space we forgo further discussion of this point here, but, for detailed consideration of the problems inherent in the attempt to ignore the alternative hypothesis, we suggest that the interested reader see Royall (1997, esp. chaps. 1 and 3) and others (e.g., Birnbaum 1962; Edwards 1972).

Suppose that, having conducted our experiment, we observe that the coin lands heads only once, on the last of seven tosses, so that $H = 1$ or $M = 7$, depending on which design we have chosen to use. Under the binomial probability distribution, the probability of this observation is $7p(1-p)^6$, so that the LR $p_2(H)/p_1(H)$ becomes $7(\frac{1}{4})(1-\frac{1}{4})^6/7(\frac{1}{2})(1-\frac{1}{2})^6 = 5.7$, favoring the hypothesis $p_2 = \frac{1}{4}$ by an LR of 5.7:1. Under the geometric distribution, this same observation has probability $(1-p)^6p$, so that the LR $p_2(H)/p_1(H)$ becomes $(1-\frac{1}{4})^6(\frac{1}{4})/(1-\frac{1}{2})^6(\frac{1}{2}) = 5.7$, again favoring the hypothesis $p_2 = \frac{1}{4}$ by an LR of 5.7:1. The law of likelihood tells us that the strength of the evidence in favor of the hypothesis $p_2 = \frac{1}{4}$ is exactly the same, once we have made our observation of one "head" in seven tosses, for, regardless of which of the two experiments we have performed, the LR is 5.7:1. Differences between the binomial and geometric probability distributions may influence our choice of experimental design prior to data collection, because the planning of statistical experiments involves the prediction of their performance across repeated enactments. But, once we have made our observation and turn our attention to the matter of what happened in this particular instance, it makes no difference which of the experiments we have chosen. The *evidence* conveyed by the data is invariant across the two experimental designs.

The *P* value does not share this invariance property. For example, if we carry out a standard one-sided test of the null hypothesis $p_1 = \frac{1}{2}$ (against the alternative, $p < \frac{1}{2}$), we find that the binomial experiment yields a *P* value of .06, whereas the geometric experiment yields a *P* value of .02. The LRs are the same under both experimental scenarios, yet the *P* values of the two tests of hypothesis differ. *Thus, when we use the* P *value as a measure of evidence, we may very well end up violating the law of likelihood.* It is also of interest that, if we performed the test of hypothesis at a predetermined significance level of $\alpha = .05$, we would not reject the null hypothesis in the first case, but we would in the second case, even though the LRs are exactly the same in each case.

The reason that the *P* values differ in the two cases is that the hypothesis-testing procedure is based on the probability distribution of the test statistic: that is, it takes into account not only the observed outcome $H = 1$ (or $M = 7$) but also the form and contents of the sample space for all other possible observations that the experiment might have produced but did not ($H = 0$, $H = 2$, . . . , $H = 7$, under the binomial design; or $M = 1$, $M = 2$, . . . , under the geometric design). The *P* values differ because the probability distribution of all the outcomes that did *not* occur differs between the two experiments. (For example, in the binomial case, there are eight possible outcomes, seven of which did not occur; but, in the geometric case, there are infinitely many possible outcomes that did not occur.) By contrast, the law of likelihood is sometimes said to imply the "irrelevance of the sample space," since no use is made of the probabilities of observations that could have occurred but did not. The law of likelihood dictates that, once an observation has been made, the strength of evidence depends solely on the observed data.

Is the "Irrelevance of the Sample Space" a Good Thing?

The law of likelihood entails the irrelevance of the sample space to evaluation of the strength of evidence conveyed by a given set of data, but is this desirable? Let us turn the question around: Why would we want the observation of one "head" in seven tosses to contribute a different quantity of evidence concerning the probability that the *coin* lands heads, depending on the *investigator's* intentions with respect to the length of the experiment? Suppose that the choice of experimental design—binomial versus geometric—had been determined by the roll of a die. For example, suppose that it had been decided beforehand that, if the die showed an even number, we would use the binomial design but that, if it showed an odd number, we would use the geometric design. Would we consider the behavior of the *die* as relevant evidence regarding the propensity of the *coin?* Surely not, and, appropriately, the probability of the observed number of dots on the die would cancel out of the LR, leaving unaffected the measure of evidence regarding the coin.

As a technical aside, we note that many frequentists, too, will agree that the behavior of the die ought to be irrelevant to the outcome of a test that purports to measure the propensities of the coin; in our simple example, conditioning the test of hypothesis on the "ancillary" outcome of the roll of the die can be used to ensure this result (Stuart and Ord 1991, sec. 31.4–31.20). But this technique is of questionable utility to much of applied statistics, and, in any event, the *P* value remains a function of a sample space, one that is defined by a single probability distribution (the "null" distribution), rather than by an LR, and it is therefore in violation of the law of likelihood. Thus the fundamental point of the example remains the same, even if, in some cases, the frequentist can rig a suitable degree of invariance for the *P* value (for further discussion, see, in addition to Royall, Birnbaum 1962; Edwards 1972).

When It comes to LRs, How Big Is Big Enough?

To return to our coin-tossing example, suppose now that we agree to use the LR as our measure of the strength of the evidence. Then we will certainly want to know how it is calibrated. How do we decide whether an LR of 5.7:1 is "strong" evidence in favor of $p_2 = \frac{1}{4}$ versus $p_1 = \frac{1}{2}$? Royall offers a "canonical experiment" to assist us here. Suppose that we are confronted with two urns, one containing all white balls and the other containing half white balls and half black balls. One of the urns is selected at random, and we randomly select (with replacement) balls from that urn. Now suppose that we observe only white balls. We can calculate the LR favoring the hypothesis that the selected urn has 100% white balls (i.e., H_1: $p = 1$, where p represents the probability of a white ball), versus the hypothesis that it has only 50% white balls (H_2:

$p = \frac{1}{2}$), associated with any run of n consecutive white balls. Specifically, if we draw b white balls, then the LR is 2^b in favor of the hypothesis that the urn contains all white balls. Thus, if we see three white balls in a row, the LR is 8; by the time that we draw 5 white balls in a row, the LR is 32; etc. Using this model as a metric, we see that the LR of 5.7:1 that we obtained in our example corresponds to drawing 2 or 3 consecutive white balls in Royall's canonical urn experiment.

Probably most of us would consider this to be relatively weak evidence regarding which urn has been selected. Accordingly, Royall suggests that we might also want to consider an LR of 5.7:1 to be relatively weak evidence in favor of $p_2 = \frac{1}{4}$ in the coin-tossing example that we have given above, since it corresponds to the amount of evidence yielded by a consecutive run of only 2 or 3 white balls in this urn experiment. Note, however, that, had we conducted a geometric experiment and performed the classical Neyman-Pearson test at the .05 significance level, we would have succeeded in rejecting the null hypothesis $p_1 = \frac{1}{2}$. This illustrates a second important but insufficiently appreciated point regarding the use of *P* values as measures of evidence: *it is possible to reject a hypothesis in the frequentist framework, or to obtain empirical* P *values falling below our significance threshold, when the actual strength of evidence, as measured by the LR, is not strong at all.*

Evidentialism in a Nutshell

Our simple coin-tossing example illustrates the evidentialist's answers to the questions with which we began this essay, as follows: Error rates of tests and measures of evidence are two different matters. The *P* value is a legitimate predictive error rate for a statistical decision-making procedure, but, as a measure of evidence, it has several undesirable properties. Chief among these is its dependence on the sample space, which is important in the planning stage prior to data collection but which should become irrelevant to the measurement of evidence once the data have been collected. The LR, in contrast, provides a sensible and straightforward measure of evidence, and, once we adopt the LR as our standard, we see that reliance on the *P* value may lead us to reject the null hypothesis when the evidence against it is weak, to fail to reject the null hypothesis when the evidence is actually strong, or, on the basis of wholly tangential differences in experimental design, to reject the null hypothesis in one case while failing to reject it in another. The evidentialist's position is, in short, that the *P* value is *not* a suitable choice for a measure of evidence—and that the LR is.

Evidentialism Meets Genetics. I. The Problem of Multiple Testing

Returning now to the debate over genomewide versus pointwise significance levels, we can see the form of the evidentialist's solution: if we agree that the *P* value, which was not invented as a measure of evidence, should not be so interpreted in scientific practice, then it follows that the "irrelevance of the sample space" renders the entire debate moot. The evidentialist views the LR in favor of (or against) linkage to any one point in the genome as the only proper measure of evidence, and this means that extraneous factors that do not affect the LR, such as the number of additional loci being tested, have no bearing on the strength of the evidence. Thus the evidentialist will say that the appearance of an unresolvable dilemma comes from treating the *P* value as if it were a valid measure of evidence—and that the solution is simply to recognize that it is not. We note that Witte et al. (1996), in their discussion of this matter, allude to the evidentialist's solution and also raise a related problem, which we have not considered here—namely, the unresolvable question of how sample size affects the interpretation of the *P* value considered as a measure of evidence. For an overview of the bewildering statistical literature on this point, see pages 70–71 of Royall.

The evidentialist's attitude toward the problem of multiple testing seems wholly appropriate to the spirit of science, for how could it be right that those who favor efficient study designs in which many different hypotheses are evaluated at once should be punished with a higher burden of proof than is required of those who examine only one hypothesis at a time? The evidentialist's solution avoids this inevitable yet irrational implication of using *P* values to measure evidence, by detaching the measurement of evidence from the often arbitrary determination of how many hypotheses are in fact being tested.

Of course, abandoning the *P* value as a measure of evidence also requires the researcher to forgo the opportunity to declare "statistical significance." But is this a problem? It has become a commonplace to include with every report of statistically "significant" linkage the caveat that a definitive finding requires independent replication and, ultimately, the cloning and functional characterization of the actual gene(s). Thus, when we say we are "rejecting" the null hypothesis (no linkage), we do not really mean that we intend henceforth to act as if the alternative (linkage) were true; what we mean is that, on the basis of the data at hand, we are inclined to accept the evidence as favoring the hypothesis of linkage. This inclination can be better expressed by use of the LR, which gives us a direct quantitative measure of the *strength* of that evidence, a measure that does not depend on extraneous aspects of experimental design.

Evidentialism Meets Genetics. II. The Problem With Model-Free Linkage Tests

Geneticists will recognize the LR in its familiar incarnation as the LOD score, which is the logarithm of the ratio of the likelihoods (i.e., LR) corresponding to the hypotheses of linkage and no linkage. By contrast, the various test statistics used by the increasingly popular "model-free" linkage tests, such as the several varieties of affected-sib-pair test (Ott 1991) or the NPL (Whittemore and Halpern 1994; Kruglyak et al. 1996), are not LRs. Because these statistics have probability distributions that can be characterized under the "null" hypothesis of no linkage, they are suitable for use in frequentist tests of hypothesis, in which the result of the test procedure is, generally speaking, calculation of a *P* value. (The probability distribution is necessary in order to take account of all those observations that could have occurred but did not.)

However, we have seen that there is reason to reject the *P* value as a valid measure of evidence in favor of the LR. It follows from this that, whatever the merits of the various model-free statistics may be, from an evidentialist perspective they are not suitable to the task of measuring the strength of evidence conveyed by our data. (Note that we are using the term "LOD score" very loosely here, without committing ourselves to a particular *form* for the constituent likelihoods: for example, the likelihood may be written under the assumption of homogeneity or by allowing for heterogeneity via an admixture parameter [Smith 1963]; it may be parameterized in terms of the usual penetrance matrices or via the recurrence risks to relatives [Risch 1990] or in terms of allelic risks [Camp 1997], etc. Thus, when we refer to a LOD score, we are really referring to a class of LR statistics. Which form of the likelihood function is optimal in any given situation is an important ongoing area of investigation. We also note that Whittemore [1996] has shown that the familiar model-free linkage statistics can be derived from the likelihood as score statistics, so that they may all be viewed as in some sense equally likelihood based; nonetheless, they are not LRs.)

Of course, a LOD score, too, can be—and often is—used as the basis for a frequentist test of hypothesis. This is possible because we can obtain a probability distribution for all possible values of the LOD score under the null hypothesis of no linkage, either by appealing to its established asymptotic relationship to the χ^2 distribution or via simulation of its empirical distribution. But, unlike the model-free statistics, because LOD scores are LRs, they also can be used in an entirely different manner—namely, as direct measures of the strength of evidence. The current debate over the relative merits of "parametric" (LOD) versus model-free linkage methods has tended to gloss over this fundamental distinction between the two approaches: the LOD score (defined broadly, as above) is not simply one among the many available test statistics; it may be the *only* one of them that is suitable to address the question, What is the strength of the evidence for linkage?

Failure to make a clear distinction between frequentist hypothesis testing and evidentialist measurement of evidence has given rise to a body of literature in human genetics in which frequentist methods are freely mixed with evidentialist objectives—a body of literature in which the *P* value is treated as a valid answer to the evidentialist's question and in which the LOD score is used to address the frequentist's concerns. The result is that we now enjoy a canon of statistical practices for linkage studies that draw simultaneously from logically incompatible first principles. The appearance of *Statistical Evidence* on the scene at this time is therefore especially timely for the field of human genetics. Readers intrigued by our brief excursion into evidentialist thinking will be richly rewarded by reading Royall's book for themselves.

And Now to the Book Itself . . .

Royall's book is written in such a way that it can be read by nonstatisticians having only a basic familiarity with the principles of statistical inference. The many useful examples rely on simple discrete distributions, such as the canonical urn experiment given above, and, occasionally, on the normal distribution, which will be familiar enough to most readers. For example, even those who are unfamiliar with the algebra of the normal distribution will appreciate the discussion (p. 50ff.) of how standard calculations systematically underestimate the required sample size. The accessibility of this section may indeed mask the fact that it represents one of Royall's unique mathematical contributions to evidentialism: the quantification of probabilities of weak or misleading evidence in strictly evidentialist terms. (This important theme is returned to and developed more fully in chap. 4; see esp. sec. 4.3ff.) Excellent (although mathematically challenging) exercises are found at the end of each chapter. Since solutions to the problems are not included, these exercises may be of limited utility; but solutions may become available in the future (R. Royall, personal communication).

Royall is able to draw on the work of some illustrious predecessors, including earlier books by Hacking (1965) and Edwards (1972; also, for the research literature in this area, see Royall's excellent bibliography). Edwards's already well-known book on this subject presupposes more mathematical background on the part of the reader, although its examples are more directly aimed at geneticists. However, Edwards is somewhat more explicit in defining the term "likelihood" itself (Edwards 1972, pp. 9–12). Readers unaccustomed to working with likelihoods might profit from reading pages 8–12 of Edwards' book, which offer a concise introduction to the likelihood and the LR (also, Ott [1991] gives an equally lucid introduction to likelihoods). Hacking's book is a somewhat lesser-known precursor. Like Royall, Hacking is primarily concerned with the logical and philosophical underpinnings of evidentialism, but nonphilosophers will probably find Royall's treatment of the subject more accessible than Hacking's.

Chapter 1, containing several loosely related subsections, introduces the important themes that occupy later chapters of the book and is vital to Royall's line of argument. The subsequent chapters each stand alone, and the reader does not necessarily need to take them in order. Chapter 2 dissects the Neyman-Pearson paradigm from an evidentialist perspective, and chapter 3 takes on Fisherian versions of frequentism. Although the breakdown of disparate points of view within the frequentist school (Neyman-Pearson vs. two distinct tendencies within the Fisherian camp) is illuminating, some connoisseurs of Fisher's work might dispute whether Fisher himself is wholly Fisherian in Royall's characterization. In particular, Royall touches only briefly on Fisher's rather abstract conception of the relevant "reference set" with respect to which the *P* value should be calculated (e.g., see Dawid 1991).

Chapter 4 pulls together the first three chapters, in a side-by-side comparison of the Neyman-Pearson, Fisherian, and likelihood (evidentialist) paradigms. Chapter 5, "Resolving the Paradoxes from the Old Paradigms," may be of special interest to clinical investigators who follow with dismay the current statistical debates for which there seems to be no possible resolution, such as the debate over the reporting of genome-wide versus pointwise significance levels, with which we began this essay (and other, equally perplexing topics, such as the question of "peeking" at preliminary data in drug studies).

Chapters 6–8 presuppose greater statistical background. Chapter 6 uses several different data sets to illustrate purely evidentialist data analyses, in comparison with standard fre-

quentist analyses. The results in each case are generally in agreement (because of the central role that the LR plays in frequentist theory as well), but the evidentialist analyses offer simplicity and nice graphical representations, along with the considerable appeal of philosophical consistency. Chapter 7 covers the variety of likelihood techniques available for the handling of nuisance parameters (marginal, conditional, estimated, profile, and synthetic conditional likelihoods). Chapter 8 contains a critique of Bayesian statistics, focusing on the inherent difficulties of relying on subjective prior probability distributions. (There is also an Appendix offering a solution to the Paradox of the Ravens, for afficcionados of the foundations of inductive reasoning.)

One small quibble is that, although Royall appears to see no valid basis to Bayesianism, he is not nearly so thorough in his criticism of the Bayes approach as he is in his critique of the frequentists. This is perhaps in part because there is greater affinity between evidentialism and Bayesianism to begin with. In particular, Royall's critique focuses on the "subjectivity" of Bayesian priors, without mention, for example, of "empirical Bayes" approaches, in which prior densities are themselves (partially) estimated from the data (Robbins 1956). Genetic linkage studies appear to offer a rare opportunity to incorporate "objective" prior distributions into data analyses, since a great deal is known about the behavior of the recombination fraction across the genome (e.g., see Elston and Lange 1975; also, for some Bayesian applications, see Smith 1959; Hauser and Boehnke 1993). It is hard to construct a scientific rationale for disallowing the use of these empirically based prior distributions. But, if they are allowed, does the frequentist testing paradigm gain validity, or are there uniquely evidentialist applications of the resulting posterior probabilities—for example, via the Bayes factor (Kass and Raftery 1995)? It would be interesting to hear Royall's opinion here.

In a similar vein, we would have been interested in greater discussion of the assessment of the strength of evidence in multivariate contexts, or in the presence of additional "degrees of freedom." It is well known that, all other things being equal, the more parameters that we estimate from the data, the larger our resulting likelihood will be. Therefore, the magnitude of the LR is affected by the difference in the number of parameters estimated in the numerator and denominator. In the frequentist approach, this difference is reflected in the degrees of freedom of the associated χ^2 statistic for nested hypotheses (Stuart and Ord 1991), but, in the evidentialist framework, there is no analogous adjustment to the LR itself. Should we interpret the strength of the evidence, as conveyed by an LR of a given magnitude, in light of the difference in "degrees of freedom?" Or does the law of likelihood tell us that all LRs of, say, 5.7:1 are created equal? Royall clearly advocates this latter position, when he points out that an LR of 4:1 represents a fourfold increase in the prior-probability ratio, regardless of the values of the priors and regardless of the context (p. 12). Thus it is fairly clear that Royall would take the strength of evidence at face value rather than attempt to make any "corrections" for the extra parameters. This seems unobjectionable in evidentialist terms, but many readers may balk at this implication of the law of likelihood, and the book gives the issue little attention. This point arises in a genetic context in the interpretation of "MMLS" (Greenberg 1989) or "mod" scores (Clerget-Darpoux et al. 1986), in which multiple parameters may be estimated in the numerator of the LOD score (e.g., see Hodge et al. 1997). The problem also arises in a slightly different form when we are expressly concerned with selection of the "best" model in a segregation analysis. For instance, some discussion of the Akaike (1973) information criterion certainly would have been interesting. But the text is, after all, a slim 176 pages and can hardly have been expected to cover all possible topics at equal length.

In Conclusion, . . .

Although the likelihood paradigm has been around for some time, Royall's distinctive voice, combined with his contribution of several novel lines of argument, has given new impetus to a school of statistical thought deserving the renewed attention of the human genetics community. For, among all the possible purposes of statistical inference, surely the measurement of evidence is first and foremost among our needs. Royall, already well-known in statistical circles for his earlier work on principles of inference and finite-population inference, has now provided a valuable manifesto for statisticians who wish to practice not just mathematics—but science.

VERONICA J. VIELAND[1] AND SUSAN E. HODGE[2]
[1]*Departments of Preventive Medicine, Division of Biostatistics, and Psychiatry, University of Iowa College of Medicine, Iowa City: and* [2]*Division of Clinical-Genetic Epidemiology, New York State Psychiatric Institute; Department of Psychiatry, Columbia University School of Physicians & Surgeons; and Division of Biostatistics, Columbia University School of Public Health, New York*

References

Akaike H (1992 [1973]) Information theory and an extension of the maximum likelihood principle. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics. Vol 1: Foundations and basic theory. Springer-Verlag, New York, pp 610–624

Birnbaum A (1962) On the foundations of statistical inference, J Am Stat Assoc 53:259–326

Camp NJ (1997) Genomewide transmission/disequilibrium testing—consideration of the genotypic relative risks at disease loci. Am J Hum Genet 61:1424–1430

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42: 393–399

Curtis D (1996) Letter to the editor. Nat Genet 12:356–357

Dawid AP (1991) Fisherian inference in likelihood and prequential frames of reference. J R Stat Soc Ser B 53:79–109

Edwards AWF (1972) Likelihood. Cambridge University Press, London

Elston RC, Lange K (1975) The prior probability of autosomal linkage. Ann Hum Genet 38:341–350

Greenberg DA (1989) Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 34:480–486

Hacking I (1965) Logic of statistical inference. Cambridge University Press, New York

Hauser ER, Boehnke M (1993) The posterior probability of linkage. Am J Hum Genet Suppl 53:1012

Hodge SE, Abreu PC, Greenberg DA (1997) Magnitude of type I error when single-locus linkage analysis is maximized over models: a simulation study. Am J Hum Genet 60:217–227

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363
Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
——— (1996) Letter to the editor. Nat Genet 12:357-358
Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318
Ott J (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore
Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241
Robbins H (1956) An empirical Bayes approach to statistics. In: Neyman J (ed) Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley
Smith CAB (1959) Some comments on the statistical methods used in linkage investigations. Am J Hum Genet 11:289–304
——— (1963) Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 27:175–182
Stuart A, Ord JK (1991) Kendall 's advanced theory of statistics. Vol 2, 5th ed. Oxford University Press, New York
Thomson G (1994) Identifying complex disease genes: progress and paradigms. Nat Genet 8:108–110
Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704–716
Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127
Witte JS, Elston RC, Schork NJ (1996) Letter to the editor. Nat Genet 12:355–356

0002-9297/98/6301-0044$02.00

Am. J. Hum. Genet. 63:289–290, 1998

Culture, Kinship and Genes: Towards Cross-Cultural Genetics. Edited by Angus Clarke and Evelyn Parsons. New York: St. Martin's Press, 1997. Pp. 272. $69.95.

This book consists of a collection of papers, from a 1994 conference in Wales, written by social scientists and health professionals interested in the impact of genetics on different cultural groups. A slim volume, it reads easily and is convenient for dipping into at leisure. There are many useful references at the end of each chapter, enabling further reading if desired. The book introduces clinical, biological, anthropological, social, and political ideas on the issues surrounding culture and kinship, with respect to genetics and counseling; it is the first publication to unite such a diverse spectrum of perspectives.

The book challenges many common Western assumptions about culture and the ethics of medical care. Unfortunately, the messages relevant to the clinician are often accompanied by complex—and, at times, angry—academic discussions. Culture is discussed mostly with reference to different ethnic groups in the United Kingdom and Africa, although the debates relevant to the United Kingdom could easily apply to any Western society. There is a brief mention of cultural groups that have formed as a result of social circumstances (e.g., people with learning disabilities could be termed a "cultural grouping"), but there is no mention of other types of culture, such as "deaf culture" or "gay culture," apart from one reference, in the editor's introduction, to deafness. One of the most poignant themes of the book is that, because rational thought is independent of race and class, clinicians who confront psychosocial difficulties in patients from different ethnic groups need to discount cultural stereotypes. Such clinicians need to acquire an understanding and respect for patients from different cultural backgrounds; the practical result of such an approach is that the patients' problems are not automatically assumed to be related to their culture.

The book opens with an insight into the different types of kinship patterns used by ethnic groups in the United Kingdom today (the European "egocentric" kinship and the Mediterranean, patrilineal, and Afro-Caribbean kinships), providing a sound basis for subsequent chapters. The discussion then turns to the various definitions of terms such as "culture," "ethnicity," "race," "society," and "relatedness." As author Helen Macbeth contends, "it is ironic that this discipline [anthropology] is centrally concerned with something that it fails to define adequately" (p. 54). Although these semantic issues are never settled, many other themes emerge as the debate continues throughout the rest of the book.

Many authors show that, although consanguinity is often blamed for disease, the networks that develop within consanguineous families can be useful for genetic counselors. There is practical information on how services for consanguineous families can be set up for the benefit of patients. Authors Sue Proctor and Iain Smith demonstrate that, despite the effect of consanguinity on increasing the risk of certain genetic conditions, it was not the main factor associated with adverse birth outcome for babies from 1,500 consanguineous Pakistani parents in Bradford, U.K. Other, more prominent factors included the mother being unable to speak English and subsequently not being directly involved in antenatal care, a problem that could be avoided if more health professionals spoke Asian languages and were better attuned to Asian culture. This example is representative of many other circumstances where the quality of care could be improved by improving cultural awareness within the medical community.

This book gives a fascinating account of how culture influences the perception of genetic disorders in the black population of southern Africa. Authors Jennifer Kromberg and Trefor Jenkins identify interesting cultural phenomena, such as belief in fate, which leads to the view that, if a child with a genetic condition is to be born, the situation cannot be altered, even by accepting prenatal diagnosis and selectively terminating the pregnancy. The authors also suggest that, when mothers want to learn *why* their child is disabled, they will consult the traditional healer rather than the Western clinician. Another point, related to language, is that there are no words for "gene" or "chromosome" in local Bantu languages. This issue also arises in genetic counseling for deafness, in which the same sign-language term for "genetics" is sometimes wrongly used also to describe "gene," "chromosome," and "DNA."

Practical issues for pedigree taking within Africa are highlighted, such as the problems encountered when the names of relationships in families are unexpected; for instance, the client's mother's older sister may be called the client's older mother, instead of aunt, or her younger sister may be called a younger mother. This particular situation also can be found