

Michael C. Acree

The Myth of Statistical Inference



Springer

The Myth of Statistical Inference

Michael C. Acree

The Myth of Statistical Inference



Michael C. Acree
Senior Statistician (Retired)
University of California, San Francisco
Cookeville, TN, USA

ISBN 978-3-030-73256-1 ISBN 978-3-030-73257-8 (eBook)
<https://doi.org/10.1007/978-3-030-73257-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gwerbestrasse 11, 6330 Cham, Switzerland

Preface

Some journeys commence without our knowing how long they will last nor where we will end up. Life, if it is done properly, is that kind of journey, and so was this book. Naturally it all seems inevitable to retrospect, though the place where I ended up will appear to almost everyone as exotic in the extreme. Yet each conclusion put me just a step beyond some other thinker, all of whom unaccountably held back from that step.

George Spencer Brown (1972), one of my favorite thinkers, was concerned to try to understand the same phenomenon in the creation of his book *Laws of Form*:

Discoveries of any great moment in mathematics and other disciplines, once they are discovered, are seen to be extremely simple and obvious, and make everybody, including their discoverer, appear foolish for not having discovered them before. It is all too often forgotten that the ancient symbol for the prenascence of the world is a fool, and that foolishness, being a divine state, is not a condition to be either proud or ashamed of.

Unfortunately we find systems of education today which have departed so far from the plain truth, that they now teach us to be proud of what we know and ashamed of ignorance. This is doubly corrupt. It is corrupt not only because pride is in itself a mortal sin, but also because to teach pride in knowledge is to put up an effective barrier against any advance upon what is already known, since it makes one ashamed to look beyond the bonds imposed by one's ignorance. . . .

We may note in this connection that Peirce [*Collected Papers*, Vol IV, Cambridge, Massachusetts, 1933], who discovered, some thirty years ahead of Sheffer, that the logic of propositions could be done with one constant, did not publish this discovery, although its importance must have been evident to him; that Stamm, who himself discovered and published [*Monatshefte für Mathematik und Physik*, 22 (1911) 137–149] this fact two years before Sheffer, omits, in his paper, to make a simple and obvious substitution which would have put his claim beyond doubt, and that Sheffer [*Trans. Amer. Math. Soc.*, 14, 1913, 481–488], who ignores Stamm's paper, is currently credited with the major discovery recorded in it (pp. 109–111).

Statistical inference, unlike the domain of abstract mathematics that Spencer Brown was exploring, is an established concept, used in daily practice by many thousands for most of a century. It, and the dualistic concept of probability which made it possible, has also been subjected to constant criticism for most of that time. If these considerations sharpen the question of why the steps I've taken here were not taken

before, perhaps they also contain part of the answer. The question, in any event, cannot be dodged if progress is to be made. Any readers who don't like where I've ended up can show me where relaxing the old constraints leads them.

Medieval cartographers put a legend at the edge of their maps, in *Terra Incognita*: "There be dragons." We laugh, thinking we know better. The dragons, however, are real, and vicious, even if they are familiar: the shrill voices of social disapproval driving us back from the edge of the already known world—for our own protection, of course. Against the metadragons.

Cookeville, TN, USA

Michael C. Acree

Acknowledgments

For nearly 40 years, I have heard what an awful experience it was to publish a book. That message came from books themselves (e.g., *The Awful Truth about Publishing: Why They Always Reject Your Manuscript—and What You Can Do about It*; Boswell, 1986) as well as from highly published colleagues. The latter also advised me of a very disturbing trend in the last couple of decades: Because nobody under 65 can any longer read well enough to enjoy it, there has been a powerful downward pressure on the length of printed books. I was assured that I would have to cut my manuscript by more than half to find a publisher. The stripped carcass that would have been left would have been unattractive, and senseless, enough to drive me to self-publish.

Nothing had remotely prepared me for the gracious acceptance and support I found at Springer, first in my content editor Sharon Panulla in New York, and no less in my production editor in Chennai, Olivia Ramya Chitrangan. Thank you both so much for making it such an unbelievably pleasant experience.

Reference

- Boswell, J. (1986). *The awful truth about publishing: Why they always reject your manuscript—and what you can do about it*. New York: Warner Books.

Contents

1	Synopsis, by Way of Autobiography	1
1.1	The Problem as I Originally Confronted it	4
1.1.1	The Place of Statistics in the Social Sciences	4
1.1.2	Statistical Inference and its Misconceptions	7
1.1.3	The Root of the Problem in the Dualistic Concept of Probability	9
1.2	How this Book Came about	12
1.3	Some Qualifications, Objections, and Implications	14
1.3.1	Intervening Developments	18
1.3.2	Some Qualifications Regarding the Historical Argument	20
1.3.3	<i>An Ad Hominem Ipsum</i>	23
	References	28
2	The Philosophical and Cultural Context for the Emergence of Probability and Statistical Inference	33
2.1	Brief Excursus on Historical Cognitive Change	35
2.1.1	Ancient Greece	35
2.1.2	Medieval Europe	37
2.1.3	General Observations	39
2.2	The Concept of Skeuomorphosis	41
2.2.1	Early Technologies of Distance	42
2.2.2	The Emergence of the Market Economy	43
2.2.3	Mathematical, Mechanistic, and Relativistic Thinking	44
2.2.4	Scientific Objectivity	46
2.3	Some Consequences for Epistemology	47
2.3.1	The Concept of Sign	47
2.3.2	The Combinatorics of Language and Thought	49
2.3.3	Causality	51
2.3.4	Representation	55
2.4	Some Psychological and Cultural Considerations	58

2.4.1	The Emergence of Self-Consciousness	58
2.4.2	Polarizations, Alignments, and Projections	60
2.5	Summary and Preview	72
	References	73
3	Origin of the Modern Concept of Probability	79
3.1	Why Gambling per se Didn't Lead to a Theory of Mathematical Probability	79
3.2	The Concept of Probability before the Seventeenth Century	82
3.3	The Calculus of Expectation	86
3.4	Statistics	91
3.5	The Art of Conjecturing	95
3.5.1	The Law of Large Numbers	96
3.5.2	The Metaphysical Status of Probability	99
3.5.3	One Word, Two Scales	101
3.6	Implications for Future Developments	107
	References	114
4	The Classical Theory of Statistical Inference	117
4.1	Bayes	118
4.2	Laplace	124
4.3	Criticism	126
	References	129
5	Nineteenth-Century Developments in Statistics	131
5.1	Descriptive Statistics	131
5.1.1	The Normal Curve in Astronomy	132
5.1.2	Application of the Normal Distribution to Populations: Quetelet	134
5.1.3	Galton, Pearson, and the Biometricians	139
5.2	Precursors of Significance Testing	143
5.2.1	Arbuthnot's and Gavarret's Use of the Binomial	143
5.2.2	Probabilistic Criteria for the Rejection of Discordant Observations in Astronomy	146
5.2.3	The Normal Model and Data in the Social Sciences	149
5.2.4	The Pun on <i>Significance</i>	150
5.2.5	Small Datasets in Agricultural Research	153
	References	156
6	The Frequency Theory of Probability	159
6.1	The Principle of Indifference	161
6.2	The Frequency Theorists	164
6.2.1	Venn	165
6.2.2	Peirce	166
6.2.3	Richard von Mises	168
6.2.4	Reichenbach	172
6.2.5	Popper	176

6.3	The Concept of Randomness	178
6.4	Summary	183
	References	184
7	The Fisher and Neyman-Pearson Theories of Statistical Inference	187
7.1	Fisher	187
7.2	The Fisherian Theory of Statistical Inference	191
7.2.1	Maximum Likelihood	192
7.2.2	Significance Testing	194
7.2.3	Fiducial Probability	203
7.3	Neyman and Pearson	209
7.3.1	Hypothesis Testing	211
7.3.2	Confidence Intervals	222
7.4	Differences Between the Fisher and Neyman-Pearson Theories	226
	References	231
8	Bayesian Theories of Probability and Statistical Inference	237
8.1	Logical Theories of Probability	238
8.1.1	Keynes	238
8.1.2	Jeffreys	244
8.1.3	Jaynes	248
8.2	Personalist Theories of Probability	250
8.2.1	Ramsey	251
8.2.2	De Finetti	253
8.2.3	Savage	256
8.2.4	Wald's Decision Theory	258
8.3	General Structure of Bayesian Inference	260
8.4	Criticism	266
8.4.1	The Logical Allocation of Prior Probabilities	266
8.4.2	The Subjective Allocation of Prior Probabilities	269
8.4.3	Subjectivity	272
8.5	Putting Theories to Work	276
	References	277
9	Statistical Inference in Psychological and Medical Research	281
9.1	Psychological Measurement	281
9.2	From Large-Sample to Small-Sample Theory in Psychology	294
9.2.1	The Concept of Probability	298
9.2.2	Significance Versus Confidence	300
9.2.3	Power	301
9.2.4	Random Sampling	301
9.3	To the Present	302
9.4	The Context of Use	305
9.4.1	Contexts of Discovery Versus Verification	305

9.4.2	Statistical Significance as an Indicator of Research Quality	309
9.5	Problems in Application of Statistical Inference to Psychological Research	310
9.5.1	Epistemic Versus Behavioral Orientation	310
9.5.2	The Literalness of Acceptance	311
9.5.3	The Individual Versus the Aggregate	312
9.5.4	The Paradox of Precision	314
9.5.5	Identification with the Null	316
9.5.6	Random Sampling from Hypothetical Infinite Populations	317
9.5.7	Assumption Violation	319
9.5.8	The Impact of the Preceding Problems	321
9.6	Possible Responses By Frequentists and Bayesians	321
9.7	Toward Resolution: The Frequentists Versus the Bayesians	323
9.8	The Recent Integration of Bayesian Concepts and Methods Into Psychological and Medical Research	326
9.8.1	Hierarchical Models	326
9.8.2	Multiple Imputation of Missing Data	327
9.9	Postscript on Statistics in Medicine	328
	References	329
10	Recent Work in Probability and Inference	335
10.1	Statistical and Nonstatistical Inference	335
10.1.1	The Putative Philosophical Distinction	335
10.1.2	Psychological Research on Reasoning in Statistical Contexts	339
10.1.3	Models of Inference	339
10.1.4	General Issues	347
10.1.5	Randomness, Representativeness, and Replication in Psychological Research	351
10.2	The Ontogenesis of Probability	354
10.3	The Propensity Theory of Probability	358
10.4	The Likelihood Theory of Statistical Inference	362
10.5	Shafer's Theory of Belief Functions	365
10.6	Recent Work on Reasoning in Philosophy and Artificial Intelligence	371
10.6.1	Some Recent Concepts from Artificial Intelligence	372
10.6.2	Bayesian versus Dempster-Shafer Formalisms	375
10.7	On Formalization	379
10.7.1	Limits	379
10.7.2	The Relation Between Philosophy and Psychology: Bayesian Theory	381
10.7.3	Purposes	383

10.8	Summary of Challenges to Bayesian Theory.....	385
10.9	Postscript on Bayesian Neuropsychology	386
	References.....	387
11	Conclusions and the Future of Psychological Research	393
11.1	Conclusions	393
11.1.1	The Concept of Probability	393
11.1.2	The Concept of Statistical Inference	396
11.2	The Future of Psychological Research	400
11.2.1	Surface Obstacles to Change	400
11.2.2	Possible Paths for Quantitative Research.....	405
11.2.3	The Ambivalent Promise of Qualitative Methods	408
11.2.4	Toward Deeper Obstacles to Change	411
11.3	The Philosophical, Social, and Psychological Context for the Emergence of a New Epistemology	413
11.3.1	Empty Self	415
11.3.2	Empty World.....	420
11.3.3	Objectivity, Skeuomorphosis, and the Problem of Scale.....	425
11.3.4	The Scarecrows of Relativism and Anarchy.....	429
11.4	Biomedical Research Without a Biomedical Model	431
11.5	Postscript on Perceptual Control Theory	434
	References.....	437
	Index.....	445

About the Author

Michael C. Acree went to graduate school at Clark University with the intention of becoming a psychotherapist, but before long it became apparent that he was too mixed up himself to be mucking around in other people's lives. Although he completed clinical training, he meanwhile became interested in the way psychologists use statistics because it didn't make any sense. He undertook a theoretical-historical dissertation to pursue the question, and the conclusion he came to after 3 years was that the reason it was hard to understand is that it literally did not make any sense. But by that point that was what he was an expert in, and could get a job teaching. But when he began trying to teach statistics to students in California schools of professional psychology, he found he wasn't that much into torturing people, and had to become a real statistician, at the University of California, San Francisco. So he figures that the major lesson he has learned in life is not to spend too much time trying to understand things that don't make sense, or that is what you'll get stuck doing for the rest of your life.

Chapter 1

Synopsis, by Way of Autobiography



To anyone acquainted with psychological or medical research, it is obvious that there is no need for yet another book—or even an article—criticizing significance testing. Devastating attacks have been published for over 50 years, and, like most critical works, they have seldom been read and have had notably little impact. This book, which started out simply as an attempt to understand that strange practice, will not dispute most of the standard complaints. But even in the best of the previous treatments, the analysis has remained at a level of superficiality that is surprising after all these years (how often these days does one get to work on a book for almost 50 years without being scooped?), mostly seeing the problems as of merely a technical nature—to be addressed, for example, by using confidence intervals in place of significance tests.

As I persisted in my efforts at understanding over several decades, the problems came to seem much deeper. The concept of statistical inference (which includes significance testing and confidence intervals, among other devices) was made possible, as I shall shortly elaborate, by the modern dualistic concept of probability, comprising an aleatory, mathematical aspect, and an epistemic aspect, having to do with knowledge, inference, and belief. It was a marriage that I don't think could ever be saved, but its powerful promise was irresistible: the formalization and quantification of uncertain inference and, in the idea of significance testing, an automatic, impersonal criterion of knowledge, making possible in turn the idea that science could be done without thought—and we suddenly got a ton of that. The same criterion conveniently applied to the evaluation of scientists: Significant results, being required for publication, and hence for grants and tenure, became an automatic, impersonal criterion for employment throughout the profession. Science could now be sponsored and carried out on a mass scale that would have been

impossible if discriminating judgment had constantly to be exercised, about theories or scientists.¹

In reframing uncertain inference, the concept of statistical inference also supported the development of a whole new concept of knowledge as something more like data than understanding. Henceforth wherever random aggregates could be found or—much more commonly—imagined, cataloguing their properties squeezed out attempts at understanding of mechanism and process, leaving us with prodigious amounts of data, and conclusions pertaining at best only to random aggregates. Psychological research has usually been safely irrelevant to real life, but the consequences in medicine are more unfortunate than we have yet realized.

The conclusions of this book are accordingly radical: that, in all the oceans of statistical bathwater flooding the social sciences, there isn't so much as a baby alligator to be saved. I could not have imagined, when I started trying to understand the use of statistics in psychology, as a graduate student over 50 years ago, that they were where I would end up. In the common conception—which may be just a stereotype, with its own supporting interests—maturation is supposed to be a process of giving up on the exuberance and excess of youth and settling down into more moderate views. My own trajectory has consistently been the reverse. I have always marveled at the ease with which, it seemed, everyone around me understood and accepted prevailing cultural beliefs and practices that made no sense to me—in science, sex, politics, or religion. (There was a time, of course, when I resented them that ease.) The more I thought about any of these areas, the less sense the established views made. An observer could be forgiven the impression that deviance was the goal.

The radicalness of the conclusions, however, imposes a burden on the exposition. What are taken to be orthodox, commonsense views have to be made to seem absurd, while the radical view has to be made to seem obvious and commonsensical. I shall make a triple, albeit very uneven, use of history for that purpose. Apart from its necessity for the purpose of understanding, and its convenience as a principle of organization, tracing the history of received views, a major task of the book, may help to undermine those views just in allowing them to be less taken for granted and in raising the possibility of alternative historical paths. Practically every field, including history itself, and mathematics most of all, is taught, on the beginning level, as a static body of “fact,” without source or direction (cf. Hadden, 1994). Samuel Goldberg, my favorite mathematics professor at Oberlin, used to lament that textbooks (not just in mathematics, I would add) are written so as to convey the impression that the currently prevailing views had always been understood and

¹The resulting factory model of science also fit beautifully with the American democratic ideal of egalitarianism. A student of mine in a Ph. D. program in psychology, protesting the difficulty of my research methods course, once claimed that “anybody should be able to get a Ph.D.” I asked if anybody should be able to get a Ph.D. in physics. “Well, no, physics is hard.” It seemed clear that graduate work in psychology was supposed to consist of gut courses, and he was feeling cheated. The idea that the most advanced degree in the field might require more work or ability than he could bring to bear was beyond consideration.

accepted; we never saw the political struggles that went into their victory or the forgotten dead ends. (See also Grattan-Guinness, 2004.) The history of science, which, as a field of study, has, at least until recently, been kept strictly separate from science itself, allows us to see how the very dilemmas and questions we presently occupy ourselves with, as advanced students, arose and formed the “epigenetic landscape” (Waddington, 1940) for succeeding generations of theories. The history of probability and statistics will prove to be a clear case in point, in giving us a valuable perspective on a set of assumptions and techniques that we now take for granted.

The second use of history will be a very brief look, in Chap. 10, at the ontogenesis of probability concepts, as a reminder of the struggles we all presumably went through with them, in trying to accommodate the curious constructions that were given to us.

The third history, which I will sketch in this chapter, is my own. Indeed, phenomenology, and not merely narcissism, would have authors introduce themselves, and identify where they are coming from, that readers might better interpret their message. The author’s self-description is of course not the final word, but a limited perspective like any other—a point to which I shall return. Especially when the message is a radical one, however, it may help to ease the entry, for those interested to undertake it, into an alien trial view, retracing how someone found the way before. Though details of the confession will be unique, I suspect that the genesis of this book, retrospectively reconstructed, taps enough common themes that it may contribute something to an understanding of how statistical inference developed and especially of why it persists.²

This first-person introduction is a late reframing, and I confess a little nervousness, just because it looks self-indulgent, and I seldom see it done.³ But part of the message of the book is about the powerful thrust toward the depersonalization of knowledge over the last three centuries (which is part of a more general thrust toward the depersonalization of authority, as manifested also, e.g., in the transition from monarchy to democracy). As Michael Polanyi said in the preface to his *Personal Knowledge* (Polanyi, 1962), “All affirmations published in this book are my own personal commitments; they claim this, and no more than this, for themselves” (p. viii). I may often speak as though my own perspective were universal, but maybe that ambition will someday come to seem both strange and destructive. To the extent that I succeed in showing how the world looks to me, in any event, others may then come to share that vision. Or not. Self-transcendence, I’m coming to think, can perhaps take care of itself in due course; we’ve been trying to rush it for many centuries now, without having noticed how poorly our attempted shortcuts were working (Koestler, 1967).

²Piaget (1950; Wartofsky, 1971) would argue that my ontogenesis should exhibit a parallel with history; but, as far as I can see, that is true only in the trivial sense that any of us acquires the current culturally constructed concepts of science before ultimately discarding them. Any such parallel in my own case would be realized only if this book were wildly more successful than I expect.

³Chris Knight’s (1991) extremely interesting work is a notable exception.

Giving away the punch line at the beginning, as I am doing in this chapter, not only violates my sense of drama; it also leaves wild conclusions hanging queasily without the support of arguments to be developed in the rest of the book. On the other hand, I wouldn't want to be accused of trying to conceal the foul pit that lies at the bottom of the slippery slope. But before we can begin our descent, I need to characterize the place of statistics in psychological research, to clarify a little further what is meant by *statistical inference*, and to identify, from a contemporary perspective, what the problem is that I originally confronted.

1.1 The Problem as I Originally Confronted it

1.1.1 *The Place of Statistics in the Social Sciences*

*Thou shalt not sit with statisticians,
Nor commit a social science.*

It was over 70 years ago that W. H. Auden (1966, p. 225) issued this commandment in a Harvard commencement address; and, at least since J. W. Tankard (1984) quoted this Phi Beta Kappa poem in his book *The Statistical Pioneers*, few critics of statistics in the social sciences have been able to resist repeating it. Grounds for Auden's concern were already evident: At around the same time, Maurice Kendall (1942), commenting on the growing influence of statisticians, observed that, "Beginning as a small sect concerned only with political economy, they have already overrun every branch of science with a rapidity of conquest rivaled only by Attila, Mohammed, and the Colorado beetle" (p. 69). Ironically, and remarkably, Kendall made his observation just before the explosive growth of statistics following World War II. With the advent of electronic computers and space-age grants to research and higher education, his comparison has long since faded into meaninglessness; with respect not only to their speed of conquest but equally to the breadth of their empire, the social scientists have put Attila to shame. In 1942, when Kendall was writing, only 29% of medical schools taught statistics, but by 1952, that figure was 82% (Marks, 1997, p. 137, n. 5). And medicine was the field where conquest by statistics was most vigorously contested.

Beginning just before World War II, the new theory of statistical inference developed during the 1920s and 1930s transformed the social sciences and neighboring disciplines. Although, with respect to issues of application, I shall be focusing on psychology, the area of my training and experience, and secondarily on medicine, the fields of education, sociology, agriculture, and economics were similarly infected. Comparative experiments, designed to investigate questions of whether certain treatments had an "effect" or not, were now assumed to have, in statistical significance, an objective criterion for yes-or-no answers to those questions, and research in these fields, in its modern forms, mushroomed in the 1940s and 1950s partly as a result. A course in statistics, with an overwhelming emphasis on

inferential over descriptive techniques, is today the one requirement across virtually all universities in these disciplines. Statistical inference has thus been regarded, almost universally for the last 70 years, as the backbone of research in the social sciences. Its monopoly is ominously secured, and attested, by the fact that it now crucially shapes the research questions that can be asked in these fields, and few practitioners can any longer conceive of alternatives.

That statistical inference should have enjoyed such splendid success is interesting inasmuch as it has been subjected for most of that time to largely unanswerable criticism (e.g., Bakan, 1966; Jacob Cohen, 1990; Lykken, 1968; Meehl, 1978; Morrison & Henkel, 1969; Oakes, 1986; Pollard & Richardson, 1987; Rozeboom, 1960). Though much of the early criticism came from biologists (e.g., Hogben, 1957), sociologists (e.g., Selvin, 1957), and philosophers (e.g., Hacking, 1965), it eventually made its way into the mainstream of psychology. The articles by Jacob Cohen (1990, 1994) and Rosnow and Rosenthal (1989) are good examples, as is Michael Oakes' *Statistical Inference: A Commentary for the Social and Behavioural Sciences* (Oakes, 1986). Both the older and newer literatures have also been widely anthologized (e.g., Harlow, Mulaik, & Steiger, 1997; Kazdin, 1992; Kirk, 1972; Lieberman, 1971; Morrison & Henkel, 1970; Steger, 1971), and the American Psychological Association some years ago appointed a Task Force on Statistical Inference (Wilkinson & the Task Force on Statistical Inference, 1999). As one result, paradoxically, it has become commonplace for psychologists to observe how little they are affected by these devastating criticisms of their fundamental beliefs (e.g., Dar, Serlin, & Omer, 1994). That ultimately became, in fact, one of the major questions motivating this essay: why we cling so desperately to a practice that is so thoroughly discredited.

One obvious reason is just the simple, generic fact that old ways tend to persist until an alternative is perceived as better. In the case of statistical inference, however, a uniquely embarrassing reason for its persistence is just that psychologists have fundamentally misunderstood it and have been accordingly reluctant to surrender their illusion. Oakes (1986) has provided a dramatic demonstration of the very widespread confusion about statistical inference, and his results will be useful here in allowing readers to assess their own understanding at the outset. (Any general readers innocent of statistical training can simply enjoy the spectacle of doing probably as well on his test as academic professionals.) Oakes asked 70 academic psychologists about the meaning of a significant result:

Suppose you have a treatment which you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means *t* test and your result is $t = 2.7$, d.f. = 18 [sic], $p = 0.01$. Please mark each of the statements below as “true” or “false.”

- (i) You have absolutely disproved the null hypothesis (that there is no difference between the population means).
- (ii) You have found the probability of the null hypothesis being true.
- (iii) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

- (iv) You can deduce the probability of the experimental hypothesis being true.
- (v) You know, if you decided to reject the null hypothesis, the probability that you are making the wrong decision.
- (vi) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions. (Oakes, 1986, p. 79).

The results are displayed in Table 1.1. The mean number of *true* responses was 2.5; only three psychologists correctly marked all six statements as false, and one of these reported having interpreted significance levels as in Statement (v) until taking this quiz and apprehending thereupon his mistake. Asked to generate any additional statements to describe how they usually interpreted significant results, only eight came up with the correct formulation, the probability of the data given that the null hypothesis is true, but five of these had also endorsed one or more of the false descriptors.

Oakes's findings appear to be robust: They have been replicated half a dozen times in the 30 years since, in as many countries; the figures have not improved. Gigerenzer (2018) has summarized them in an important paper.

The pervasive confusion about statistical inference goes far toward explaining its ritualistic use. Indeed, what enabled the extreme rapidity of the takeover of the social sciences by statistical inference was in part that, unlike most religious practices, it essentially bypassed the step of meaningful practice and went straight to ritual. The prominence of costume metaphors among its critics is one indication that it functions not so much as a research tool as a spiritual, or at least a psychosocial or political, tool, as it were. Social scientists, Suzanne Langer (1967) observes, appear to believe "that the dress of mathematics bestows scientific dignity no matter how or where it is worn" (p. 39). But these fancy clothes often have no emperor, no substantial body of theory underneath. Leamer's (1983) observations about economics are no less applicable to psychology:

Table 1.1 Frequencies of *true* responses to interpretations of significance tests

Statement	f	%
(i) the null hypothesis is absolutely disproved	1	1.4
(ii) the probability of the null hypothesis has been found	25	35.7
(iii) the experimental hypothesis is absolutely proved	4	5.7
(iv) the probability of the experimental hypothesis can be deduced	46	65.7
(v) the probability that the decision taken is wrong is known	60	85.7
(vi) A replication has a 0.99 probability of being significant	42	60.0

Note. From *Statistical Inference: A Commentary for the Social and Behavioural Sciences* (p. 80) by Michael Oakes, 1986, Chichester: Wiley. Copyright 1986 by M. Oakes

Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our *t*-values. (p. 37)

1.1.2 *Statistical Inference and its Misconceptions*

With Oakes's pretest out of the way, it is time to concretize the initial focus of this essay by providing an informal clarification of what is meant by statistical inference and taking a brief look at some of the criticisms. For simplicity I shall be speaking in this chapter only of significance testing; other techniques comprised by statistical inference (e.g., confidence intervals) I shall get to in due course.

In a typical significance test, we might want to determine whether our treatment had had an effect. Now it might be supposed that we could just look at our data and see whether there were a difference between our treatment and control data. But when psychologists speak of a treatment effect, they are talking about a difference that is unobservable in principle. We actually imagine our treatment and control data to have been randomly sampled from a hypothetical infinite population of such scores, and our statement of an effect pertains to a difference between these hypothetical populations. When we do a significance test, we start by assuming something we typically don't believe for a moment—for example, that the treatment has no effect. The significance level we calculate is the probability of drawing two such different samples if there were no difference in the population means. If this probability lies beyond a conventional cut-off, we reject the null hypothesis of no effect.

So the first thing that is strange about the procedure is that it is wholly hypothetical. Except in survey research, we don't take random samples, but that is the basis for our probability calculations, which otherwise have no meaning.

The second thing that is odd about this procedure is that we are calculating the probability of the data, which we already have in hand, rather than the probability of the hypothesis. Significance levels are often misinterpreted as the probability that the null hypothesis is true; but the probability of the data, given the hypothesis, is not the same as the probability of the hypothesis, given the data. It is the difference between the probability that someone is Catholic, given that he is the pope, and the probability that someone is the pope, given that he is Catholic.

The argument also has a peculiarly triple-negative logic. As Lancelot Hogben (1957, p. 337) put it, speaking of the laboratory research worker: "For what reason should he or she be eager to take advantage of a procedure which can merely assign a low probability to erroneously asserting that the treatment is useless?"⁴

Even earlier, Berkson (1942) commented on the peculiar logic of tests which are designed only to nullify a negation: "With the *corpus delicti* in front of you, you do

⁴Hogben's use of nonsexist language—in speaking of laboratory research workers—6 years before *The Feminine Mystique* (Friedan, 1963) is remarkable (though perhaps a little less so if we consider the issues faced by someone named Lancelot).

not say, ‘Here is evidence against the hypothesis that no one is dead.’ You say, ‘Evidently someone has been murdered’” (p. 326).

Fourth, notice that the process culminates in a decision respecting the hypothesis, rather than an evaluation. As Rozeboom (1960) said, we treat our hypotheses like a piece of pie offered for dessert—accept or reject, “Thanks” or “No, thanks.” That’s a strange way to treat scientific hypotheses, and in fact it is not what we do in practice; but it’s what the theory says we are supposed to do.

Finally (for the moment), the probability we calculate refers to the long-run relative frequency of false rejections of true null hypotheses. But in scientific research, we don’t care about the long-run average performance of a testing process; we care about each individual hypothesis. Again, significance levels are commonly misunderstood as the long-run proportion of our decisions, or perhaps of our rejections, which are correct; but in fact we would never have any idea what proportion of our decisions were correct unless we knew whether the hypotheses we were testing were true; and if we knew that we wouldn’t be testing them.

The pervasive misconceptions implicit in these criticisms, and documented by Oakes (1986), have been addressed in the psychological literature—Jacob Cohen (1990), pointing obliquely to inadequacies of statistical instruction, subtitled the first section of his article “Some Things You Learn Aren’t So”—and there are signs that the criticisms are beginning to have discernible effects. Over the last 50 years, for example, textbooks of statistics for psychologists have become noticeably more accurate in their characterization of significance testing and confidence intervals and more cautious about their ritualistic use. Correspondingly, journal articles are no longer so likely to report naked p values, without information relevant to effect sizes. Considerable as they are, these changes would still have to be regarded as modest for a span of over half a century. Even if, however, contrary to present trends, statistical inference within the next few decades should pass into the archives of psychological research methodology without ever having been generally understood in the first place, thus limiting the present book largely to chronicling a curious sidebar of history, the issues go well beyond psychological research. Although only psychology and neighboring disciplines have made statistical inference a formal methodology, the concept and its theoretical underpinnings reach deeply into not only the philosophy and psychology of knowledge but also into the modern lay understanding of cognition. Since the issues are complex, it is difficult to convey in this introductory overview a full sense of what the problems are, but the synopsis in the following section at least offers an idea of what lies ahead and will also make possible a more specific statement about the contribution of this book.

1.1.3 *The Root of the Problem in the Dualistic Concept of Probability*

The source of confusion over statistical inference is the concept of probability, on which statistical inference rests. It is evident from reflection on everyday usage that the concept of probability covers several aspects or meanings. On the one hand, it advertises to frequency of occurrence in some set or series of events. If we say “set,” we have the classical, or Laplacean, measure of probability from gambling as the ratio of the number of cases favorable to an event to the total number of equiprobable cases; if the definition is to avoid circularity, judgments of equiprobability must appeal to some principle of symmetry or indifference.

If we say “series,” we have the frequency definition of probability, in terms of a sequence of events, whether real or ideally constructed. The frequency definition is often confused with the classical, but the distinction is important. The frequency conception is regarded by its proponents as having the advantage, over the classical, of being empirical; actuarial probabilities based on mortality statistics are the cardinal example. The classical ratio, in contrast, is based on an *a priori* model (e.g., of a fair coin). Either definition presupposes a random process (rolling a die, deaths from the plague), with the probability of an event either assumed or estimated from data; calculations then concern probabilities of particular outcomes of the random process.

Probability, in addition to being conceived as a relative frequency, can also refer to a degree of expectation or belief. Here again the interpretations divide. The logical version takes probability to be a degree of *rational* belief, or the degree to which a body of evidence warrants a particular assertion. The personalist theory, also known as the subjectivist theory, purports to take probability as an individual’s degree of belief, though it imposes constraints, like additivity (The probabilities of a proposition and of its denial must add to 1.), on any system of beliefs, or belief coefficients, to make them behave like classical probabilities. Probability on these epistemic theories is not limited to random processes but pertains to uncertainty from whatever source. There are other interpretations, too; a more recent one takes probability to be a physical propensity of a device or set-up to display a certain outcome.

It can be, and has been, argued, for any of these meanings, that it can comprehend all instances of the use of *probability*; but most examples, rather than fitting all interpretations equally well, can be understood much more easily and naturally as one meaning or the other. If we ask, for example, about the probability of a royal flush at draw poker, it seems clear enough that we are referring to classical probability, and there would be some question about taking the exact value 1/649,740 as the measure of our expectation of that event. If, on the other hand, we ask about the probability that peace in the Middle East will be achieved within the next 10 years, we seem to refer to the epistemic meaning, and a frequency interpretation, though possible, is more strained.

Whatever our particular formal theory of probability, these two aspects, by this point in history, are given in our experience, almost like the construction of space as three-dimensional. We can see the pull of the epistemic aspect simply by trying to imagine that an event is improbable—peace in the Middle East, for example—while divorcing ourselves from the expectation that it will not happen. We can see the givenness of the aleatory aspect in the fact that, at least for anyone capable of formal operations, probability automatically maps onto a 0-to-1 scale corresponding to relative frequencies. The mapping may be implicit in particular instances; but, if pressed, we would all assign 1 as the upper limit of probability and 0 as the lower, whatever these values might mean; and we would find it unnatural to think of the value, say, -1 as representing the lower limit of probability for such an event or proposition.⁵

The frequency and credibility theories diverge crucially on the issue of probabilities of individual cases: probabilistic statements about whether it will rain tonight, whether there is intelligent life in this galaxy, whether the next toss of a given coin will turn up heads, whether the trillionth decimal of π is a 9, whether there really is no difference between the means of the respective populations from which the experimental and control groups in a given study were sampled, and so on. Theories which interpret probability in terms of degrees of reasonable belief, subjective betting odds, or the like have no special trouble with examples of this kind; but the frequency theories face the problem of definition of the reference class. It is usually possible to *invent* classes in which to embed a given event; but to maintain a claim of objectivity, these theories need to be able to argue that these classes are at least well defined and unambiguous, if they are not always obvious or unique. The problem does not appear to admit a general solution.

Hypotheses are ordinarily regarded as a special kind of individual case. Mathematical statisticians sometimes work with sequences of hypotheses, as in defining asymptotic relative efficiency for nonparametric tests; and some probability theorists have defined probability by reference to the relative truth frequency of hypotheses in some appropriate sequence; but, as scientists, we are nearly always interested in *particular* hypotheses. The truth of our hypotheses, moreover, is generally not determined by a random process, either random sampling or random assignment. On both of these counts, therefore, hypotheses can plausibly be assigned probabilities only on an epistemic, not on a frequency, interpretation of probability.

But here arises the fundamental dilemma. The received theory of statistical inference—the Neyman-Pearson theory—was designed expressly to accommodate the frequency interpretation of probability,⁶ which restricts the theory to “direct” probability statements, such as statements of statistical significance: probabilities of outcomes on some model or given some hypothesis. The so-called Bayesian approach adopts an epistemic orientation to probability, taking the fundamental meaning as a

⁵ Or maybe not: Shortliffe and Buchanan (1975) defined a “certainty factor” for use in their MYCIN model of medical reasoning, for which the values -1 and $+1$ represent, respectively, the limits of disconfirmation and of confirmation.

⁶ On Neyman’s equivocation, see Chap. 7.

degree of credibility or rational belief. The latter conception, in contrast to the frequency view, allows “inverse” probability statements to be made: statements about hypotheses (or propositions or events) given some data.

As many sources (e.g., Bakan, 1967; Carver, 1978; Oakes, 1986; Pollard & Richardson, 1987), if not the standard textbooks, explain, many of the expressions commonly used to describe significance levels—“the probability that the results are due to chance,” “the probability of the null hypothesis being true,” “the degree of confidence warranted by the null hypothesis,” “confidence in the repeatability of the results”—involve an implicit inversion of significance levels, which is illegitimate in the orthodox theory. “The probability that the results are due to chance” is but another way of saying “the probability that the null hypothesis is true.” No such locution is possible in the Neyman-Pearson theory.

The Bayesian approach thus poses a serious challenge with its strong *prima facie* claim to greater relevance. For insofar as the purpose of statistics in psychology is construed as evaluation of degree of support awarded hypotheses by data, or modification of opinion in light of observation, or the like, the Bayesians are alone in even claiming to satisfy it. The traditional theory, on the other hand, terminates in a yes-no decision respecting hypotheses, and its probability statements concern the long-run frequency of errors in such a decision process. It has found its least controversial application in quality control in manufacturing, where repetitive sampling is a reality and accept-reject decisions are themselves what is needed. The Bayesian approach can boast additionally the sly advantage of being able to duplicate frequentist results by suitable choice of prior belief function (see Chap. 8).

The problem has been obscured for psychologists by the fact that their textbooks have propagated a confusion which they took over from Ronald Fisher, encouraging the epistemic interpretation of frequency probabilities. The orthodox textbook methodology is itself an uneasy, anonymous blend of concepts from two schools of notorious mutual hostility, the Fisher and Neyman-Pearson theories. Students in statistics courses in the social sciences, despite their complaints, have actually not realized, in some ways, what a cruel position they were in. The controversies discussed here have mostly been hidden from them, and the attempts of textbook writers to paper over the differences—or, worse, their failure to understand the issues—have been partly responsible for the genuine difficulty of such courses. Students learn (more or less well) the proper things to say about statistical inference, twisting themselves to accommodate a perverse reality, but there is very little opportunity for most to achieve a real understanding. Many books have been written to make statistical inference intelligible, presenting it in ever simpler terms; but the problem with most of them is that they speak from within the orthodoxy and present merely more forced rationalizations of it.

1.2 How this Book Came about

This is more or less where I came in. The main thing that was unusual about my experience was just my discomfort with all the confusion. This book really started with my embarrassment at how much trouble I had in understanding statistics, in my own first-year graduate course at Clark University. It wasn't the instructor, Neil Rankin, or the book, William Hays' (1963) classic text (then in its first edition); both were the best that I have encountered. Nor was it my preparation: As an undergraduate mathematics major, I should have been in an unusually good position to understand the material. But I found it hard to give a coherent exposition of significance testing when I was assisting in teaching undergraduates and started reading further.

So the original goal was simply understanding, with no expectation of ever having an original thought. But as Poul Anderson is famously supposed to have said: "I have yet to see any problem, however complicated, which when you look at it in the right way did not become still more complicated."⁷ His observation would suggest that I have a disarmingly good record of looking at problems in the right way. In the present case, once I got to (what I now take to be) the root of the problem, the growing complications appeared as inevitable consequences of an incoherent framework within which the problem was posed, and the existing critiques (continuing up to the present) seemed hopelessly—not to say, safely—superficial.

My starting point was, however—almost of necessity, if it were truly a starting point—conventional enough. "Good student" that I was, I had conscientiously absorbed the frequentist doctrine when I was first exposed to it as an undergraduate psychology major.⁸ By the time I began my own serious study in the area, however, I was expecting to strengthen my incipient impression that Bayesianism made more sense. Since the distinction between the Bayesian and the orthodox theories pivots on their concepts of probability, I thought, in a thoroughly rationalistic way—I was in my early 20s—that I would get to the bottom of the dispute with a conceptual analysis of probability. I was not the first to be intrigued by that inscrutable concept; there are many book-length treatments of the subject. The developmental orientation at Clark University gave me at least a primitive framework for approaching problems, by trying to understand how they got that way. Up to that time, history had meant to me memorizing the governors of Arkansas (At least one of my classmates can still recite the whole string, a skill which gets pitifully little exercise outside class reunions.) and the details of Civil War battles. It was essentially *military* history, as though people had never done anything but try to kill each other, and I had avoided it as much as possible. I had had little exposure to the history of

⁷The origin of the quote is not easy to trace; I am coming to suspect that Anderson never wrote it down. There are references to an article by William Thorpe, "Reductionism vs. Organicism," in the *New Scientist*, September 25, 1969. But the *New Scientist* supposedly started publication in 1971. Never having been a science fiction fan, I encountered the quotation in Arthur Koestler's *The Ghost in the Machine*, published 2 years earlier, in 1967.

⁸I was hardly alone. Maurice Kendall (1949) admits: "Myself when young did eagerly frequent" (p. 103).

science or to intellectual history more generally—to the development of the ideas that much of the fighting was about—which turned out to be much more interesting.

A few years after I began work in this area, Ian Hacking published *The Emergence of Probability*, which filled in some gaps more solidly than I would have hoped to. I was intrigued by a passing reference to work by Glenn Shafer, and the following year Shafer published his dissertation as *A Mathematical Theory of Evidence* (1976). Shafer's theory now did for Bayesian theory what Bayesian theory had done for frequentism, namely, to cannibalize it: As Bayesian theory can duplicate frequentist results by imposing suitable restrictions, Bayesian theory itself dropped out as a special, somewhat bizarre, case of Shafer's theory of belief functions, which accommodated everyday usage still more adequately than Bayesianism. I was persuaded by Shafer (even if he wasn't) that what were called subjective probabilities—like the probability of a hypothesis being true or of our entering another ice age within the next century—were not really subject to meaningful measurement and calculation like frequency probabilities.

Shortly after the appearance of *A Mathematical Theory of Evidence*, Shafer published an enormously important historical paper, “Non-additive probabilities in the work of Bernoulli and Lambert” (1978), in which he argued that it was largely a matter of accident that the term *probability*, formerly a nonmathematical concept meaning something like trustworthiness, got attached to the mathematical concept of chances which emerged in Europe in the seventeenth century. The curiously dualistic concept that resulted, tying degrees of evidence or belief with mathematical calculation, immediately held out the promise of solving the skeptical problem of induction, through the concept of statistical inference.

By its very meaning, statistical inference is obliged to comprehend both aspects of the concept of probability: To qualify as inference, it appeals to the epistemic meaning; to qualify as statistical, it appeals to the frequency, or aleatory, meaning. The dualistic concept of probability thus functions, as it were, like a detergent molecule, grabbing hold of the grease with one end and the water with the other, and carrying them both down the drain. Chap. 3 traces how these concepts, chance and inference, came to be associated in the epistemology of the Enlightenment. However strained that assimilation, it goes deep enough by now in this culture to be part of reflective lay wisdom. Michael Cowles (1989) is in this respect a useful witness, in his struggles to understand the concept. When he says, by way of justifying statistical inference, “*All* inference is probabilistic and therefore all inferential statements are statistical” (p. 27), many readers would take his statement as unremarkable and obvious; yet it clearly rests on an elision from an epistemic to an aleatory interpretation of probability.

It will nevertheless be the argument of this book that these two notions, chance and inference, are not really compatible and that the concept of statistical inference was thus never a viable one in the first place. It should perhaps have come as no surprise in my historical research that theories committed to the requirements of statistics—in particular, the orthodox Neyman-Pearson theory—should have abandoned all reference to inference, becoming instead a theory of statistical decision-making, and that theories attempting to remain faithful to the requirements of

inference, in particular, Shafer's theory of belief functions, should have been led to abandon statistics. The dualistic concept of probability has bedeviled philosophers for three centuries, but the conclusion seemed plain to me that these two concepts should never have been stuck together in the first place, and that inference and statistics really have nothing more to do with each other than, say, ethics and trigonometry.⁹ The surprise is just that Shafer himself should have shrunk from the conclusion to which his work seems so clearly to lead.

The fundamental conclusion of this essay is thus the claim, stronger than I have seen elsewhere, that the forced union of the two aspects of probability is a sterile hybrid, inspired and nourished for 300 years by a false hope of formalizing inductive reasoning, making uncertainty the object of precise calculation. Because this is not really a possible goal, statistical inference is not, cannot be, doing for us today what we imagine it is doing for us. It is for these reasons that statistical inference can be characterized as a myth.^{10,11}

1.3 Some Qualifications, Objections, and Implications

Especially in view of the radicalness of my conclusions, it will be important to clarify at the outset what I will *not* be challenging. (Maybe these count as baby alligators.) (a) In the first place, I am not directly concerned here with descriptive statistics. These presuppose for their legitimate use, however, both valid measurements and a primary interest in aggregates, and I will be questioning the

⁹ Not everyone would find the latter far-fetched. Francis Edgeworth, who contributed to the development of the correlation coefficient, wrote a book (Edgeworth, 1881) pursuing the geometrization of morals, complete with transformations to polar coordinates and all the rest, and “jealously purged,” as Langford Price said of Edgeworth's writings on economics, “of the corruption of concrete content” (quoted in Barbé, 2010, p. 131). “The application of mathematics to the world of soul is countenanced by the hypothesis . . . that Pleasure is the concomitant of Energy” (Edgeworth, 1881, p. 9). “The invisible energy of electricity is grasped by the marvelous methods of Lagrange; the invisible energy of pleasure may admit of a similar handling” (p. 13). I was startled, in particular, by Edgeworth's proposal that pleasure (measured in *jnds*) is distributed, in the plane of means and capacities, as a megisthedone. I had always vaguely thought *megisthedone* was the name of a giant precursor of the woolly mammoth—either that or the mother of all tranquilizers.

¹⁰ In calling it that, I am following the usage of writers like Szasz (1961, 1978) and Sarbin (1968), who have performed destructive analyses of the concepts of mental illness, psychotherapy, and anxiety; but I confess that this usage, which was especially popular in psychology for a while, makes me a little weary. In the first place, it seems to reduce and cheapen a concept with which I associate some grandeur and profundity. Beyond that, it connotes a hard-nosed, facile debunking of a concept that may actually designate some important if delicate truth. So, I feel some reluctance to identify myself with such a program, even though I don't think there is anything more sacred about statistical inference than Sarbin thinks there is about anxiety. Nevertheless, that concept seems to be the most widely understood for what I want to say about statistical inference.

¹¹ Oakes (1986), possibly taking a cue from Acree (1978/ 1979, p. 477), also concludes that “Statistical *inference* is a myth” (p. 145; his emphasis).

applicability of both these conditions in psychological research and the latter in medical research. (b) Among the so-called inferential procedures of psychological research, my criticism does not affect randomization tests, so long as they are applied to true experiments (but not to the analysis of intact groups; see Chap. 7) and are properly interpreted. These have historically had only marginal importance, though there is reason to expect them to play a larger role in psychological research, as software packages and finally textbooks catch up with the analyses made possible by computers. (c) I make no special quarrel with the Neyman-Pearson theory as a decision procedure for contexts of repetitive sampling where interest lies only in long-run performance of the test. Similarly, I have no objection to standard procedures for estimation of population parameters, so long as the intervals are appropriately interpreted. Hence, the application of traditional “inferential” procedures to problems like manufacturing quality control or large-scale opinion surveys stands secure. But all these are of little relevance to psychologists, who have been characteristically concerned with tests of singular hypotheses based on small samples (from which estimates would be practically useless anyway).

The message of this book will not be a welcome one nor an easy one for psychologists to assimilate. There are unfortunately at least two reasons beyond the direct threat posed by the content. One of the potential difficulties in taking its message seriously has to do with the trickiness of methodology in general as a target of criticism. As Jeffreys (1961) remarks, methodological criticism “is usually treated as captious, on the grounds that the methods criticized have delivered the goods” (p. 417). Jeffreys wants to argue that the methods he criticizes—I shall criticize the same ones, as well as his—have not, in fact, “delivered the goods”; but the argument is more subtle than his discussion indicates, because it is not apparent how we can tell whether “the goods” have been delivered or even what they are. In particular, we can easily be seduced by a large body of research findings into supposing that they provide some special vindication of significance testing. That is the embarrassingly vapid conclusion of Michael Cowles’ *Statistics in Psychology: A Historical Perspective* (1989):

The plain fact of the matter is that psychology is using a set of tools that leaves much to be desired. Some parts of the kit should perhaps be discarded, some of them, like blunt chisels, will let us down and we might be injured. But, they seem to have been doing a job. Psychology is a success. (p. 199)

Psychology is unquestionably a raging success—in terms of having sustained itself for more than half a century as a lucrative jobs program for an intellectual elite. And the practice of significance testing has played a very substantial role, in providing an automatic, impersonal criterion for judging the value of research reports and hence of their authors. But if psychology is a success, it is certainly not due to the scientific merit of statistical inference, for indeed there is no way to assess just what kind of job statistical inference has been doing. If it were utterly useless (not from a political, but from a scientific, standpoint)—if the practice of statistical inference were equivalent, let us say, to using a Ouija board to answer our research questions—then what we should expect throughout the field is random,

nonreplicable patterns of results. But, as Meehl (1978) and others have pointed out for years, that is just what we do find. It is hard to recall literature reviews which failed to conclude that the findings were mixed and that further research was therefore needed to clear everything up once and for all. Phenomena appear in the literature, generate a flurry of research, and then disappear. Of course, it has also been pointed out that the much discussed “decline effect” is limited to those social sciences which rely on statistics (http://thelastpsychiatrist.com/2011/02/the_decline_effect_is_stupid.html). Greenwald (1975) suggested years ago that at least some of these may simply be epidemics of Type I errors, and Ioannides (2005a; b) has more recently documented the same phenomenon in medicine. The problem doesn’t lack recognition; Marcia Angell (2009) famously said:

It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgment of trusted physicians or authoritative medical guidelines. I take no pleasure in this conclusion, which I reached slowly and reluctantly over my two decades as an editor of the *New England Journal of Medicine*.

Her counterpart at *The Lancet*, Richard Horton (2015), writes:

The case against science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness.

A second reason why the message of this book will be difficult to assimilate concerns the nature of critical essays in general. Strictly negative works, in themselves, tend, with some reason, to be idle and ineffective. Feyerabend (1975) reminds us that no theory of anything has ever been fully adequate, and consequently the job of the critic is the easiest in the world. Or, as Kendall (1942) puts it, “Any stigma will do to beat a dogma” (p. 71). But if even very bad theories—we shall see some examples—are not relinquished until something better comes along, the special trap with the concept of statistical inference is that there are no good theories to be had, and there never will be. Hence, to criticize statistical inference is unhappily like attacking conventional religious mythology with its promise of life everlasting: It requires a radical reorientation to imagine the world a better place without that belief. I will suggest in the concluding chapter what the reorganized fields of psychological and medical research might look like, but that vision I owe to Bill Powers (1973, 1978), and can only point to.

Given that I score 100% N on those Myers-Briggs tests, the practical question of what to do without statistical inference is one that might never have occurred to me, except that it was one of the first questions I was asked when I began lecturing on my research. It struck me as odd that psychologists would be so helpless in interpreting their data; if the data were so ambiguous that professional investigators needed the help of a statistician, I was inclined to say that they surely had their answer already.

My first take on a more elaborate answer was that Jacob Cohen (1962, 1988) had pointed the way with his work on power and effect size. Prior to his work, the concept of power had been largely neglected by psychologists, partly because it was

difficult (some early textbooks omitted it for just that reason), but more fundamentally because it pointed explicitly to the subjectivity of the procedure: It was up to the investigator to specify what effect size was of interest. Cohen, somewhat against his wishes, solved that problem by proposing conventions, analogous to the .05 and .01 significance levels. But once investigators were prepared to specify an effect size of interest, then significance levels became superfluous: They could just look and see whether that effect size obtained. As descriptive statistics, effect sizes eliminate the problems of statistical inference, but at the same time all of its appeal, in terms of rendering an automatic judgment.

So long as I continued teaching, I gave the issue pathetically little more thought—in particular, the fact that effect sizes were still descriptive *statistics*. Once I realized I didn't enjoy torturing people enough to continue trying to teach statistics to psychologists and became a real statistician, I gained a more vivid impression of how fundamentally misguided contemporary research practice was, in both psychology and medicine. The use of statistics in both disciplines is essentially epidemiological. Epidemiology is a legitimate science, looking for patterns in large aggregates that are invisible to the naked eye. But it is also essentially a propaedeutic science: Those patterns are interesting not in themselves, but in pointing to underlying causal processes of potential interest. We are also properly wary of epidemiological data, because they are associational. Most research in psychology, and much of it in medicine, is a hybrid, with epidemiological analysis of experimental data, making ritualistic reference to fictitious populations. A result of greatest consequence is that the statistical analysis has come to constitute the stopping point as well as the starting point of inquiry. Especially in psychology, but also in pharmacology, we don't get beyond the epidemiological patterns to probe the underlying processes. I am astonished at how often, when I look up drugs in the *Physicians' Desk Reference*—drugs that are typically designed to be as powerful as possible—the mechanism of action is said to be unknown. Even our inquiries into process, like mediation analyses, are still statistical. No other method is accredited.

The statistical model specifies that a treatment acts on the average of a random aggregate, with individual deviations regarded as “errors.” We know the model is false—psychological and biological processes operate within individuals, not on averages of random aggregates. What Hayek (1935/ 2008) said years ago about economic analysis is no less relevant to psychology and medicine:

Neither aggregates nor averages do act upon one another, and it will never be possible to establish necessary connections of cause and effect between them as we can between individual phenomena, individual prices, etc. I would even go so far as to assert that, from the very nature of economic theory, averages can never form a link in its reasoning. (p. 200)

Although most psychological research is inconsequential enough to protect the discipline indefinitely from methodological blunders, the implications of the statistical model are particularly clear in medicine. Andrew Vickers (2005), in many respects a fine critic himself of statistical practice, writes: “A typical question a patient might ask would be: ‘If 100 women were given adjuvant chemotherapy, how many of them would have a recurrence compared with 100 women who had only

surgery?”” (p. 652). It would take a statistician, or at least a thoroughly educated person, to think such a thing. What we all want to know, on the contrary, is whether the treatment in question will help *us* in particular. It is an indication of how far the statistical mentality has pervaded our thinking that we have implicitly given up on that question as hopeless. But, if the many billions of dollars which have gone into conventional clinical trials to determine a small average effect of a drug across hundreds or thousands of cases had gone instead into studying its biological properties and the reasons for individual differences in response, medicine might be a less scattershot affair, with doctors trying first one cholesterol or depression medication and then another. Wolfram (2020) argues that his principle of computational irreducibility “in biology is going to limit how generally effective therapies can be—and make highly personalized medicine a fundamental necessity” (p. 550). It is ironic that, just as medicine is beginning a return to an idiographic approach, with the emergence of subdisciplines like pharmacogenomics, the bureaucratization of medicine in the political sphere is pushing it powerfully back into the statistical model dealing with average patients and one-size-fits-all treatments. *So twentieth-century.*

1.3.1 Intervening Developments

It would be extraordinary if we could spend half a century chronicling the development of a very active discipline without that history being constantly rendered obsolete by changes in the field. Yet that is very nearly what happened in the writing of this book. The criticisms that were raised in the last 50 years, occasionally by prominent psychologists, were not new and had no more effect than the original ones. Within the past two decades, however, a shift has been discernible, and it parallels that occurring in response to Jacob Cohen’s work on power (1962). Up to that time, power was a neglected concept of statistical inference (Chap. 9). It was difficult and for that reason omitted from many texts; many of the textbooks which did discuss it made mistakes; and it was generally ignored in research work. Cohen suggested that under the circumstances we might expect that the power of our experiments to be low, and he estimated that perhaps half of our experiments would be capable of detecting an effect if there were one. It is interesting to ponder why scientists are so impervious to logic; but Cohen’s empirical, numerical demonstration had an impact—a very gradual one: Within 30 or 40 years, the federal government was requiring power analysis in grant proposals.

A similar logic applies to the replication problem, with the fact that *p* values are such poor arbiters of truth. It was Ioannides’ (2005a; b) quantitative estimate, similar to Cohen’s, that perhaps half of all published research findings are false that finally caused medical and psychological research workers to take notice. Within a few years, Brian Nosek (Open Science Collaboration, 2015) had instigated a quantitative, empirical demonstration of Ioannides’ claim. In a massive collaborative effort, they selected for replication 100 articles in the year 2008 from 3 premier

journals in psychology: *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. They achieved a replication, in the sense of also getting significant results, in 39. The whole fantastically expensive enterprise of psychological and medical research is thus arguably less than worthless, because it leads to reliance on findings which are known not to be reliable: better not to have produced the research in the first place.

Not content with this many-studies/one-replication demonstration, Nosek (Silberzahn et al., 2018) also spearheaded a one-study/many-replications approach, in which a single, complex dataset was sent to 29 analytic teams, with the same research question. The dataset was a realistic one: The structure was complex enough to offer several different analytic possibilities, and the data were ambiguous enough that choices of analytic strategy could make a difference in the results. The teams used 21 unique sets of covariates, though 1 author thought the inclusion of 1 particular covariate by some teams was indefensible and should be counted as a mistake. Twenty of the teams got a significant result, with widely varying effect sizes; 9 did not. That variation is pretty much what a seasoned analyst would have expected, though the article still caused quite a stir.

Predictably, there has been a new round of proposed solutions, none of them really new. One journal, *Basic and Applied Social Psychology*, did go so far as to announce, in 2015, that it would no longer accept papers reporting *p* values. But the authors of one proposal to “abandon statistical significance” (McShane et al., *in press*) are careful to note that they “have no desire to ‘ban’ *p* values” (p. 3). If history is any guide, we should expect to see in another 50 years a fresh round of awakening, hand-wringing, and feckless proposals—with no real change. But in all these developments, there are two fundamental questions that have never been asked.

1. How could so many bright people, for so many generations, have remained indifferent to such grave flaws in their methodology? Some of the answers are obvious and are mentioned in this chapter. If these add up to something less than fully satisfying, we shall be considering whether, on a more emotional, psychological level, the practices still felt right.
2. More importantly, there has never been any recognition of the fundamental way in which significance testing shaped the research questions asked in psychology. A century ago, it was not obvious that all questions in psychology should be framed in epidemiological terms. Yet none of today’s critics of significance testing can even conceive the possibility of research questions being framed in any other terms. We shall have to glimpse some such possibilities here.

The most radical proposal, unsurprisingly, has come from a nonscientist: James Corbett is a wide-ranging investigative journalist, who proposes a genuinely “open science” (www.corbettreport.com/openscience). Corbett is astute enough to realize that there is one simple, yet sufficient, explanation for psychologists’ persistent reliance on a discredited methodology: They are very well paid for it. Such an anomaly could not arise in a free market for research funding; but with the government’s near-monopoly on psychological research funding, it could, and probably will, go

on until a hyperinflation destroys the currency. Intellectuals who praise Eisenhower's prescient warning about the military-industrial complex in his Farewell Address of January 17, 1961, forget that it was but one of two threats he identified as having arisen during his administration; the other was the scientific-technological-Congressional complex (not his term):

Partly because of the huge costs involved, a government contract becomes virtually a substitute for intellectual curiosity. . . .

The prospect of domination of the nation's scholars by Federal employment, project allocations, and the power of money is ever-present—and is gravely to be regarded.

Yet, in holding scientific research and discovery in respect, as we should, we must also be alert to the equal and opposite danger that public policy could itself become the captive of a scientific-technological elite.

Hence Corbett calls for a Feyerabendian anarchist regime in science. A paradigm case of such an investigation occurred in response to the Fukushima nuclear disaster in 2011, the Safecast program. Many Japanese citizens were concerned that the government wasn't leveling with them and was even evacuating residents from lower- to higher-radiation areas. So they simply started taking their own measurements. Citizens carried Geiger counters attached to GPS devices, which were set to relay readings every 5 seconds, creating a live map of radiation levels all over the country. Eventually the government surrendered to its authority and changed the evacuation zones.¹²

A practical emergency requiring only meter readings would seem tailor-made for such a public science venture; but the concept has been applied to advanced mathematics as well. In 2009 Timothy Gowers (www.gowers.wordpress.com/2019/01/27/Is-massive-collaborative-mathematics-possible/), a Fields medalist, posted an unsolved problem on-line and invited help. Contributors built on each other's work, and in 37 days, the problem was solved, with papers subsequently published. (Neither of these examples involved the use of statistics, obviously.)

The most ambitious and impressive open science collaboration is *The Project to Find the Fundamental Theory of Physics* led by Stephen Wolfram (2020).

1.3.2 Some Qualifications Regarding the Historical Argument

The “mythical” nature of the concept of statistical inference itself virtually necessitates a historical analysis, not only of statistical inference but also of the concept of probability on which it is based. Treatises of probability—like J. R. Lucas's (1970), for example, to pick one virtually at random—which begin, as I once aspired

¹²When the project was implemented in Washington, DC, it was discovered that the World War II memorial was made of highly radioactive granite. The federal government had known, but hadn't released the data.

to, from reflection on the modern concept are doomed to a certain futility, just because they necessarily accept its crucial presuppositions and are therefore blind to them. The very deep confusion over the meaning, extensional as well as intensional, of *schizophrenia* is similarly pathognomonic, perhaps, of a mythical concept that should have died or shouldn't have been born. Like statistical inference and probability, the concept of schizophrenia survives less because of what it actually designates than what it promises; and to get at the root of these concepts entails a willingness, at least in principle, to relinquish that promise. Not by asking directly what statistical inference, or probability, *means*, in short, but rather how these concepts *came about*, can we discern whatever meaning they may be said to have. Outside that context, the orthodox theory of statistical inference, in particular, is simply too arbitrary and Procrustean a framework to make sense of.

One of two opposite objections may come to mind at this point. (If both do, there is a diagnosis for that, and welcome to the club.) One is the observation that revealing the history of an idea is not the same thing as refuting it. To show, as I do, that our favored data analytic tools derived from misguided attempts to mathematize induction does not invalidate them *per se* as data analytic tools, any more than the utility of Post-It™ notes is nullified by the failure of their adhesive as a glue. Nevertheless, this book may, if it is successful, help to bring about change in the same way that insight-oriented psychotherapy sometimes does: Once we see clearly why we have been doing certain things, it becomes harder to keep doing them.

The opposite concern is that all our concepts may be vulnerable to this kind of destructive analysis and that there is accordingly nothing special about statistical inference or schizophrenia. “All concepts are destructible,” writes Spencer Brown (1957) in his own provocative analysis of the concept of randomness, “and it is not always obvious which to destroy” (p. v). But concepts differ in their capacity for damage containment, as it were. As a more limited example, we may consider what can be done to the concept of *woman* by simple historical scrutiny. Even though English has no less offensive alternative label for a female adult human, I have always wondered at the eagerness with which feminists embraced this particular one. *Woman* is, first of all, a contraction of *wife-man*, as if no other status were conceivable for a person of that gender. But *wife* itself, coming from the Latin *vibrare*, to vibrate, originally meant one who hovers or flutters restlessly about (Partridge, 1966, p. 776). It is also related to the word *whip*—“through the German, of course” (Keys, 1972, p. 138). The obnoxious connotations are a nuisance because they are not extinct; but the specifically etymological connections are by now distant enough—as witness, the fact that the word is championed by those whom it most directly offends—that it is only cultural factors which stand in the way of a purification or redemption of the concept. The word can presumably be ready whenever we are, in other words, to strip the oppressive wrapping off the package. With statistical inference or schizophrenia, on the other hand, the problem is not just the packaging of a certain bias with a legitimate referent. These concepts take their meaning only in the context of notably inadequate theories. Hence, attempts to examine them within their own respective contexts leave them intact in their obscurity, whereas

critical scrutiny of the frameworks themselves leaves those concepts without support.

To construct a history on so large a scale as I am attempting here also involves some considerable risks, which it would be well to note in advance. As we might expect from the strangeness of the outcome whose origins we are tracing, the story is not a smooth one, studded as it is with improbable events. More than once, an unpromising hybrid seed happened to fall into a manure pile in a neighboring field and flourished luxuriantly. Nevertheless, it is still too easy to draw together small threads of history into a single strand of spurious continuity. This fallacy is aided by the ease of anachronistic interpretation of terms and themes from centuries ago. With histories of science particularly (as well as the genetic epistemology of Piaget, 1950), it has been a common temptation to construct a linear developmental sequence, with an endpoint toward which all individuals or cultures are evolving and which they will someday reach, last one turn out the lights. This kind of anachronic interpretation has become known, following Butterfield, as Whig historiography (Kragh, 1987). Obviously, however, those whose work I discuss in the seventeenth through the nineteenth centuries were not trying to provide modern psychology with a methodology for research. Contemporary statistical inference, rather than something centuries of work went into trying to develop and perfect, will turn out to be more like the byproduct of unsuccessful attempts to develop something else. I expect, nevertheless, that some of the continuity I find in this book will be controversial. A certain case could be made, for example, against taking any history of contemporary statistical inference back much farther than a century, to the work of Galton, Pearson, and Fisher. Certainly the epistemological problem that Fisher took up had seen no significant progress since the pioneering work of Bayes and Laplace a century and a half earlier, and there is a legitimate question whether, given the great conceptual and cultural developments of the intervening years, Fisher's problem was really still the same as theirs. I will argue, in fact, that it was and that the epistemological basis of the theory of contemporary statistical inference is itself singularly anachronistic.

A second issue in historical writing arises from the fact that necessarily much of the discourse herein will concern the writings of individual men. (The pronoun is intentional, and the near absence of contributions from women is not necessarily to their discredit.) Just because it relies so heavily on the written record, intellectual history, just as easily as political history, can appear to espouse, wittingly or not, a simple “great men” theory of history. If any one individual was the sine qua non of modern statistical inference (and consequently the sine qua plus ultra¹³ of psychological research), it was R. A. Fisher, with his particular mathematical genius and keen interests in both philosophy and applied research. For the rest, it is surely true that, had some of these individuals never lived, somebody else would have come up with something similar eventually. On the other hand, it is not completely apt to see this historical line as preformed and waiting to unfold, in the hands of whomever it

¹³Thank you, Lee Kaiser.

happened to fall. It is possible to imagine some different possibilities from what actually occurred, and it will be instructive to notice some of these as we go along.

A third problem in writing such a history is achieving even a passing acquaintance with all the relevant literature over a span of 400 years. Inevitably I have relied on secondary sources, especially before the seventeenth century. The occasions of such reliance are generally made clear in the text, but here I want to acknowledge a long-term risk with this practice in general. We are all familiar with the distortions of communication that occur in the party game of Telephone. It is very dismaying to see the same effects in the history of thought. As one or two historical accounts become standard for a field, they tend to become the only sources subsequently referred to, and the particular distortions imposed by their selection and interpretation become the new reality. Disciplinary histories can indeed be effective instruments of political change within a field; Mitchell Ash (1983) has described several instances of this process in textbook histories of psychology. The practice grows rapidly as few American academics can any longer read any language besides English. But I suspect anyone returning to original sources has at least occasionally had the experience of finding startling discrepancies from the standard, textbook accounts. In the present field, an outstanding example occurs with the classical work, James Bernoulli's *Ars Conjectandi* (1713). Virtually all attention to this book (my own included) had focused on the last few pages, which provided the first proof of the law of large numbers (popularly, and somewhat misleadingly, known as the law of averages). Glenn Shafer (1978) was the first person in almost 300 years to notice, just a few pages earlier, the profound implications of Bernoulli's work for the development of the concept of probability.

1.3.3 An Ad Hominem Ipsum

A final issue in the construction of history has to do with the fact that the intellectual development I have sketched so far, as well as most of the history in this book, is what could be called *conscious* history: an account of what various people *thought*. But the conscious level is not necessarily where the action is; we need to attend as well to the cultural and psychological conditions which make possible or difficult a given line of development, the usually invisible background conditions which phenomenology aims to uncover. “Unconscious” history is a tricky endeavor, however. It easily becomes reductionist, or determinist, and is ordinarily taken as *undermining*, or invalidating, the conscious arguments; psychoanalysis itself, in fact, is conceived as a “hermeneutics of suspicion” (Ricoeur, 1970). The *argumentum ad hominem* is not merely a logical fallacy; it also readily, and understandably, gives offense. But I don't know that, in undertaking such an inquiry, we are committed to the cynical view that the unconscious level necessarily holds the truth, which conscious discourse is frantically concerned to cover up. We can surely take account of factors promoting or hindering certain lines of thought without seeing them as decisive. I shall be endeavoring, at any rate, to listen with one ear for psychological and

cultural, as well as philosophical, themes in the development of probability and statistical inference. At the cultural level, for example, the question arises of why the concepts of probability and statistical inference should have appeared in England and France rather than, say, Italy or Japan. Questions of why something didn't happen can be difficult, however; I will have little to say about them except in Chap. 3, where the story begins.

It may be useful to include here a brief *ad hominem ipsum*, with illustrative examples of such factors in the genesis of the present book. A “conscious” history will tend to make it sound as though any curious person, at least in a given culture, should have ended up in the same place. However strange it may seem to me, that is clearly far from being so in my own case. Identifying a few additional influences, conditions, and supports affords also the opportunity for some appropriate acknowledgments. Some of these are “unconscious” only in the sense that I was not aware of their role at the time, and their relevance was general rather than specific to the content of this book.

Chronologically the first influence, of course, was my parents. I confess that acknowledgments of the author’s parents had always struck me as puerile and trite; it is also true that my parents, not being intellectuals, made no direct contribution to this book. But over time it has become clearer what extraordinary support I had been taking for granted. My mother, whose intelligence, in more favorable circumstances, would have allowed her to go much farther in that direction than being a crack stenographer, had an independent streak which perhaps found in me its greatest opportunity for expression. And I still marvel at how my father, with a tenth-grade education and solidly conventional views, could have managed to be so supportive of whatever I became. I would not have done nearly so well with a son whose main interests were hunting, football, and boxing. When I see how much more difficult it appears to be for many people to hold views fundamentally different from most of those around them, I think I may know one of the reasons.

A somewhat more specific, yet still very global, contribution was made by John Lau. At the time we met, starting graduate school, my thinking, if it could be called that, was terminally rigid and defended, as necessary, by well-practiced tactics of intimidation. It took an impossible level of both intellect and caring to break through that wall; I don’t know anyone else who could have done it. I had started out confident that I would convert him to my views in 6 months; I was the one who yielded, through almost daily conversations about epistemology over a period of 2 years. Whatever flexibility my thought may now claim, if not necessarily many specific conclusions, I owe to John.

The Psychology Department at Clark, with its theoretical, philosophical orientation—Continental philosophy at that—and its lack of structure at the time provided an environment that was uniquely suited to my style, if not to the views I entered with. Neil Rankin made the initial suggestion that I study the development of probability concepts for my dissertation, but that wasn’t going to yield the conceptual analysis I was looking for (and Piaget & Inhelder, 1951, had already studied the development of the aleatory concept of probability). I’m profoundly grateful to Lenny Cirillo for his suggestion that I do a theoretical dissertation and to Rachel

Falmagne for taking on the task of chairing the committee. The support from them, and from the entire Psychology faculty, was tremendous.

Since I left the nest of graduate school for the real world, virtually all of my intellectual work has been extracurricular. Given additionally my introversion, that has meant that I have always worked alone, without benefit of professional conferences or connections. Had I known anyone interested in the same issues, and knowledgeable about them, I might have been steered back on track long ago.

I wasn't literally living on a desert island for 40 years, however, and I should acknowledge a few individuals who made a serious effort to make my work known and to establish some collegial ties. In the summer of 1987, at the Pacific Graduate School of Psychology in Palo Alto, I was working on a second draft of this book, thanks to Richard Chapman and a generous grant from the Chapman Research Fund. Scanning the shelves of a Berkeley bookstore, I came across a copy of *The Probabilistic Revolution* (Krüger, Gigerenzer, & Morgan, 1987), newly published. I had not heard of many of the contributors, and so was extremely surprised to notice my name in the index. Gerd Gigerenzer had referred to my dissertation (Acree, 1978/ 1979) in his article "Probabilistic Thinking and the Fight against Subjectivity" (1987). I was even more surprised a couple of weeks later to get a call from Gigerenzer, who, it turned out, was only a couple of miles away, at the Institute for Advanced Study in the Behavioral Sciences at Stanford. He had heard of me, as I recall, through Kurt Danziger, whose work I had read and admired; and it was evidently my fellow Clark alumna Gail Hornstein who had introduced Danziger to my dissertation.¹⁴ Gigerenzer had been interested in getting in touch with me, but I was totally unknown, and it wasn't easy in the days before the Internet. The preceding year, however, he had been at Harvard, and his young daughter Thalia was at Harvard Day Care. When he was discussing his work one day with the Director, Gary Yablick, Gary said that he knew someone who would be interested in it: Gary, another Clark alumnus, was a former housemate of mine, and we had been in touch recently enough that he could put Gigerenzer in contact. I expect this might have been a better book had I accepted Gerd's exhortation to apply for a Fulbright scholarship for study in Germany. Gerd's own work has been widely read, and I have been able to trace my own influence through a couple of generations of citations.

In the following 5 years, I was grateful to Larry Roi, Christopher Lyall Morrill, Michael Scriven, and Ed Lieberman for reading and commenting on the manuscript (or portions of it) and to members of the "Dawn Seminar" at the California Institute of Integral Studies—Tanya Wilkinson, Robert Rosenbaum, William Irwin Thompson, Don Johnson, Eva Leveton, Jurgen Kremer—for critiques on the first two chapters. Gail Hornstein, my tireless publicist, also introduced me to Elihu Gerson at the Tremont Research Institute in San Francisco, who gave me the most detailed and helpful critique I received. It was natural that I, a clinical psychologist by training, should have looked to psychology for explanations of our attachment to

¹⁴It must have been Gail who, a few years earlier, also passed along my dissertation to Persi Diaconis and Arthur Dempster, both of whom were generous enough to send encouragement and suggestions.

statistical inference; Elih was very helpful in encouraging me to look more closely at cultural factors. In more recent years, after another round of revisions, Wolf Mehling and David Anderson read and commented on the early chapters. Within the last 5 years this book was in preparation, I met Michel Accad though San Francisco's Circle Rothbard. Accad, a cardiologist and Thomistic philosopher, took a dim view of medical research methodology; his book *Moving Mountains: A Socratic Challenge to the Theory and Practice of Population Medicine* (2017) is a fine critique of the concept of population as patient. Accad had also established, with Anish Koka, a weekly podcast, the Accad-Koka Report; Brian Nosek, of the Open Science Collaboration, contributed an important interview (www.accadandkoka.com/episode48). Eager to give advance publicity for my book, he invited me also to do an interview (www.accadandkoka.com/episode57). All of these—except possibly for Tanya—will be dismayed to see how little of their advice I ultimately took, but I'm grateful for all of it.

When anyone expressed interest in my book, I generally offered Chap. 2, as the least technical and most accessible to a general audience. I was startled at how sharply responses fell into two groups, generally along gender lines. Women were likely to say that they could hardly get through the first half, but that in the second half it all came together beautifully. Men tended to say that the first half was solid, but that it all fell apart in the second half. I suppose only a bisexual could have been so oblivious to the difference between the two parts or fail to be repelled by one of them.¹⁵

To return to conditions which helped to make it more comfortable for me to work alone, and to reach deviant conclusions, the most interesting is one whose recognition I owe to Lin Farley, and one which I have not seen discussed elsewhere. Most academics are in fact not in a good position to have noticed it, but I also doubt I would ever have discovered it myself.

I had not heard of Lin when she enrolled in my Social Psychology class at the California Institute of Integral Studies; she was the author of *Sexual Shakedown* (1978) and a member of the Cornell group who originated the term *sexual harassment*. Lin was a journalist, not an economist or political scientist; so I was all the more impressed with the economic and political analysis in her book—a fresh, independent look unencumbered by the usual preconceptions. Lin told a story in class of a reception at the Waldorf-Astoria held for her by McGraw-Hill on the publication of her book. Since her parents lived in New York, she expected to see them there; then she realized that, given their status as what the left likes to call the “working class,” they would have felt out of their element in the Waldorf-Astoria and would not have set foot inside. Surely enough, she found them waiting outside at the curb. I told her that I could imagine my parents having some of the same feelings, and she

¹⁵ George Spencer Brown might disagree. Although he says explicitly that his book *Only Two Can Play This Game* (Keys, 1972) was written from a feminine perspective (albeit under the decidedly masculine pseudonym of James Keys), he characterizes relationships prior to the one that book was about as having been “thin and homosexual” (p. 21) in comparison. One infers from this epithet that any gay experiences he may have had were so impoverished as to have lacked electricity.

surprised me by saying, “I thought so.” I had said nothing about my background. But after she told me about the basis of her inference, I began to notice that, with surprising reliability, I could identify which of my fellow academics came from blue-collar backgrounds. Some of these, of course, are eager to shed their outsider status and adopt the protective coloration of the intellectual environment they have moved into. For those of us who may be more indifferent, or oblivious—or hopeless—on the other hand, the distinguishing characteristic, to a first approximation, is iconoclasm; but I think a more precise specification might be an attitude of irreverence toward the hallowed customs and pretensions of academentia. I once attended a lecture by a prominent neuropsychologist who, in the course of describing an experiment with rats, said, “Sadly, we cracked open their heads and pulled out their brains.” That simple, literal utterance, frankly acknowledging what he had done, and implicitly his discomfort with it, would have been unthinkable for most of my colleagues; it elicited nervous laughter among those present. The appropriate expression, “The animals were sacrificed,” uses the noble-sounding word *sacrifice* and puts the action discreetly in the passive voice. I said to the speaker afterward, “I’m guessing that your father was not an academic.” By way of startled confirmation, he guffawed. Obviously a blue-collar background by no means guarantees that an academic will reach unconventional conclusions, but it does appear to help. For those born into the proper class, on the other hand, the socialization begins early, is largely invisible to the in-group, and makes heterodox attitudes and opinions more difficult to reach. Robert Grosseteste, an independent thinker of the thirteenth century, and founder of the science of optics, knew Paris, according to Pye (2015), but was still from provincial England.

With the “unconscious” motivation of the author in the more traditional, psychoanalytic sense, we come closest to understanding where this book is coming from. I had probably been trying for 10 years to understand the concepts of probability and statistical inference before it occurred to me (as it had all along, I’m sure, to some of my clinician friends) to wonder why the problem had seemed so urgent to me, that I would have spent so many years studying it.

When I began supervising dissertation students in clinical psychology, it didn’t take me long to realize that there was basically just one dissertation research question: namely, “Is there anybody else out there like me?”. Students typically didn’t have a research question; they just wanted to talk to other people who had experienced incest, eating disorders, alcoholic parents, or any of the other defining characteristics of clinical psychologists. I was a clinical psychology student myself once; so, how does one end up doing a dissertation in statistics? It’s a question I could have answered, interestingly, only after I had reached some conclusions—as though I could see what I was doing only after I identified ostensibly the same motives in others. I had to admit, when I thought about it, that it had a lot to do with my having been so hung up on rules. It bothered me to see people using the orthodox theory in an implicitly Bayesian way and getting away with it. Once I got around to reflecting on it, it was quickly enough apparent that I had all of my life had an overly heavy investment in rules—rules as such—and that I had been paying more of a price than I had been aware. Or, rather, I hadn’t allowed myself to

consider that the price might not have been a necessary one. It was also obvious that my interest in methodology was very much of a piece with my early adolescent preoccupation with ethics—which, sad to say, was my particularly abstracted way of trying to cope with conflicts about sex. (I suggest in Chap. 2 and again in Chap. 11 that sexual conflicts played a larger role in the genesis and establishment of statistical inference than has usually been imagined.) In any event, the solution I envisioned originally was to confront those who were inconsistent in their practice with an irrefutable demonstration of their inconsistency; then, I suppose, they would have had to be as unhappy as I was.

Long before I might have formulated such an irrefutable argument, however, I came to realize that the Truth didn't necessarily matter all that much to people. Not just to jerks, but regular people. Even for *me*, I noticed, it was a little bit of a strain to lose an argument. Winning seems to be more important than being right or even than being nice. (Robert Nozick, 1981, is unusual among philosophers in having made some similar observations.) But there lay a double defeat for statistical inference. Like logic, it is supposed to formalize argument, to be the objective, impersonal arbiter of dispute. Yet it seemed there was little reason to hope that it could serve its intended purpose. The principles of deductive logic are not commonly disputed, yet neither are they effective in settling a controversy. Rare is the argument of consequence that can be resolved, for example, by pointing to an undistributed middle term. Statistical inference shares this weakness with deductive logic and carries the additional burden of having always been the object of bitter controversy itself. Still, psychologists cling to it *as if* it uniquely filled a vital arbitrative function.

It began to appear to me that the real purpose it was serving must be something different from what is commonly pretended, and I began to think about possible such functions. The clinging to rules, to a point where science was hindered rather than advanced; the establishment of a strong, depersonalized authority; and the attitude—indeed the philosophy—of self-distrust and self-abnegation which gave rise to the felt need for mechanisms of control, all immediately suggested parallels with politics and with ethics. In the same movement, the possibility of a thoroughgoing anarchism, comprehending all these domains, presented itself. The final chapter, after exploring some of the roots of our perceived needs for control, of ourselves and others, will consider whether that idea is altogether as naïve as we would like to believe. Chapter 2, meanwhile, will explore the interesting relevance of money—and sex—in laying the epistemological groundwork for the emergence of probability and statistical inference.

References

- Accad, M. (2017). *Moving mountains: A socratic challenge to the theory and practice of population medicine*. Green Publishing.
- Acree, M. C. (1979). Theories of statistical inference in psychological research: A historico-critical study (Doctoral dissertation, Clark University, 1978). *Dissertation Abstracts International*, 39, 5037B. (University Microfilms No. 7907000).

- Angell, M. (2009, January 15). Drug companies & doctors: A story of corruption. *New York Review of Books*. Retrieved from <http://www.nybooks.com/articles/2009/01/15/drug-companies-doctors-a-story-of-corruption/>
- Ash, M. (1983). The self-presentation of a discipline: History of psychology in the United States between pedagogy and scholarship. In L. Graham, W. Lepenies, & P. Weingart (Eds.), *Functions and uses of disciplinary histories* (pp. 143–189). Dordrecht: Reidel.
- Auden, W. H. (1966). *Collected shorter poems*. New York: Random House.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437.
- Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco: Jossey-Bass.
- Barbé, L. (2010). *Francis Ysidro Edgeworth: A portrait with family and friends*. Northampton, MA: Edward Elgar.
- Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37, 325–335.
- Bernoulli, J. (1713). *Ars conjectandi* [the art of conjecturing]. Basel: Thurneysen.
- Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378–399.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cowles, M. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, NJ: Erlbaum.
- Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, 62, 75–82.
- Edgeworth, F. Y. (1881). *Mathematical psychics*. London, UK: C. K. Paul.
- Farley, L. (1978). *Sexual shakedown: The sexual harassment of women on the job*. New York: McGraw-Hill.
- Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London, UK: New Left Books.
- Friedan, B. (1963). *The feminine mystique*. New York: Norton.
- Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1, 198–218.
- Grattan-Guinness, I. (2004). The mathematics of the past: Distinguishing its history from our heritage. *Historia Mathematica*, 31, 163–185.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hadden, R. W. (1994). *On the shoulders of merchants: Exchange and the mathematical conception of nature in early modern Europe*. Albany: State University of New York Press.
- Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hayek, F. A. (2008). *Prices and production and other works* (J. T. Salerno, Ed.). Auburn, AL: Mises Institute. (Original work published 1935).
- Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart, Winston. (2nd ed., *Statistics for the social sciences*, 1973; 3rd ed., *Statistics*, 1981; 4th ed., 1988).
- Hogben, L. (1957). *Statistical theory: The relationship of probability, credibility and error*. New York: Norton.

- Horton, R. (2015). Offline: What is medicine's 5 sigma? *The Lancet*, 385, 1380. Retrieved from <http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736%2815%2960696-1.pdf>
- Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, 294, 218–228. <https://doi.org/10.1001/jama.294.2.218>
- Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon Press. (1st ed., 1939).
- Kazdin, A. E. (Ed.). (1992). *Methodological issues and strategies in clinical research*. Washington, DC: American Psychological Association.
- Kendall, M. G. (1942). On the future of statistics. *Journal of the Royal Statistical Society*, 105, 69–80.
- Kendall, M. G. (1949). Reconciliation of theories of probability. *Biometrika*, 36, 101–116.
- Keys, J. (1972). *Only two can play this game*. New York: Julian Press.
- Kirk, R. E. (Ed.). (1972). *Statistical issues: A reader for the behavioral sciences*. Monterey, CA: Brooks/Cole.
- Knight, C. (1991). *Blood relations: Menstruation and the origins of culture*. New Haven: Yale University Press.
- Koestler, A. (1967). *The ghost in the machine*. New York: Macmillan.
- Kragh, H. (1987). *An introduction to the historiography of science*. Cambridge, UK: Cambridge University Press.
- Krüger, L., Gigerenzer, G., & Morgan, M. S. (1987). *The probabilistic revolution (Vol 2: Ideas in the sciences)*. Cambridge, MA: MIT Press.
- Langer, S. K. (1967). *Mind: An essay on human feeling* (Vol. 1). Baltimore: Johns Hopkins Press.
- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73, 31–43.
- Lieberman, B. (Ed.). (1971). *Contemporary problems in statistics: A book of readings for the behavioral sciences*. New York: Oxford University Press.
- Lucas, J. R. (1970). *The concept of probability*. Oxford: Clarendon Press.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151–159.
- Marks, H. M. (1997). *The progress of experiment: Science and therapeutic reform in the United States, 1900–1990*. Cambridge, UK: Cambridge University Press.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (in press). Abandon statistical significance. *American Statistician*.
- Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology*, 46, 806–834.
- Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *American Sociologist*, 4, 131–140.
- Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy*. Chicago: Aldine.
- Nozick, R. (1981). *Philosophical explanations*. Cambridge, MA: Belknap Press.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester: Wiley.
- Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science*, 349, 943–951.
- Partridge, E. (1966). *Origins* (4th ed.). New York: Macmillan.
- Piaget, J. (1950). *Introduction à l'épistémologie génétique* [Introduction to genetic epistemology]. Paris, France: Presses Universitaires de France.
- Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hazard chez l'enfant* [the development of the child's concept of chance]. Paris, France: Presses Universitaires de France.
- Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy* (Corrected ed.). Chicago: University of Chicago Press.

- Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*, 159–163.
- Powers, W. T. (1973). *Behavior: The control of perception*. Chicago: Aldine.
- Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review, 85*, 417–435.
- Pye, M. (2015). *The edge of the world: A cultural history of the North Sea and the transformation of Europe*. New York: Pegasus Books.
- Ricoeur, P. (1970). *Freud and philosophy: An essay on interpretation*. New Haven, CT: Yale University Press.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416–428.
- Sarbin, T. R. (1968). Ontology recapitulates philology: The mythic nature of anxiety. *American Psychologist, 23*, 411–418.
- Selvin, H. C. (1957). A critique of tests of significance in survey research. *American Sociological Review, 22*, 519–527.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.
- Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences, 19*, 309–370.
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences, 23*, 351–379.
- Silberzahn, R., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*, 337–356.
- Spencer Brown, G. (1957). *Probability and scientific inference*. London, UK: Longmans, Green.
- Steger, J. A. (Ed.). (1971). *Readings in statistics for the behavioral scientist*. New York: Holt, Rinehart and Winston.
- Szasz, T. S. (1961). *The myth of mental illness*. New York: Hoeber-Harper.
- Szasz, T. S. (1978). *The myth of psychotherapy*. New York: Doubleday Anchor.
- Tankard, J. W., Jr. (1984). *The statistical pioneers*. Cambridge, MA: Schenkman.
- Vickers, A. J. (2005). Analysis of variance is easily misapplied in the analysis of randomized trials: A critique and discussion of alternative statistical approaches. *Psychosomatic Medicine, 67*, 652–655.
- Waddington, C. H. (1940). *Organisers and genes*. Cambridge, UK: Cambridge University Press.
- Wartofsky, M. W. (1971). From praxis to logos: Genetic epistemology and physics. In T. Mischel (Ed.), *Cognition and epistemology* (pp. 129–147). New York: Academic Press.
- Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594–604.
- Wolfram, S. (2020). *The project to find the fundamental theory of physics*. Wolfram Media.

Chapter 2

The Philosophical and Cultural Context for the Emergence of Probability and Statistical Inference



In an article on the origins of statistics, the quotable Maurice Kendall (1960) observed:

A history must start somewhere, but history has no beginning. It is therefore natural, or at least fashionable, for the historian of any particular period or of any particular subject to preface his main treatment by tracing the roots of his theme back into the past as far as he possibly can. The writer on any modern idea who can claim that the Chinese thought of it first in the Shang period is usually regarded as having scored a point. (p. 447)

The point is, however, apt to be stretched, as Kendall's benign sarcasm reminds us; and, in point of fact, the history of probability and statistics has not been traced as far back as the Shang period. Van Hée (1926), Sheynin (1974), and Mahalanobis (1957) have unearthed references to ideas resembling chance or likelihood in Eastern literatures, in the last case as far back as the sixth century B.C. Hacking (1975) has found a more recent (fourth century A.D.), but more interesting, Indian epic in which, with strikingly modern sophistication, estimation of the number of leaves on a tree was linked with the science of dice. As he himself contends, however, there is no reason to suppose any connection between these concepts, as interesting as they may be in their own contexts, and the development of probability in Western Europe.

More commonly, the history of probability, like that of every other Western concept, is traced to ancient Greece (e.g., Kneale, 1949; Reichenbach, 1949), and a certain ordinary and quite misleading account of the “early history of probability” can be written just by chronologically arraying references to terms that have in modern times been translated (not always, of course, by someone sensitive to the history of this concept) as *probability*. Remarkably, however, a good case can be made that the only history relevant to the present subject begins rather abruptly in the mid-seventeenth century, when the modern dualistic concept of probability appeared. The origins of that concept, and of the concept of statistical inference which it made possible, are the subject of Chaps. 3 and 4. The present chapter is concerned to describe the philosophical and cultural context of the emergence of

these concepts, especially with regard to the changing view of knowledge at that time.

The seventeenth century has been known as the age of the Scientific Revolution—of Galileo, Newton, and Boyle, the founding of the Royal Society, and so on—but in recent years a number of scholars (e.g., Berman, 1984; Bordo, 1982/1983, 1987; Easlea, 1981; Foucault, 1973b; Keller, 1985; Merchant, 1980) have taken a new look at that era in Western Europe and have documented some features hitherto less noticed. It was a time of profound epistemological upheaval—I say “upheaval” because it appears to have had the character of a sudden orogeny deforming the landscape and destroying familiar routes and landmarks, as much as it had the character of a planned revolt against earlier ways. Philosophically, the overall results of the change included the naturalization of authority, a shift in the metaphysical status of language, the disappearance of Aristotelian causes and a general externalization of knowledge, and the formation of sharp splits between subject and object, mind and body, science and religion, public and private, masculine and feminine.

These changes may have been gradual over the lifespan of any individual; but, on a historical scale, they were surely abrupt enough to have made themselves felt, even at an individual level. Though historians and philosophers are only beginning to pursue the Collingwoodian¹ question of what it must have felt like to live during that time, when we see, in the following centuries, the earnest eagerness with which silly “solutions” were seized upon for the new problems engendered by the loss of connection, I think we at least get a clue to the answer. Psychologically, the overall effect of the changes can be summarized as an alienation from nature, a loss of a sense of belonging and of place—above all, for our interests, the loss of the old certainties. Minor (for the present purpose), but still very significant, themes were the rejection of the past, of childhood, of origins, starting anew; and, paralleling the separation between self and world, an estrangement between the sexes and a renewed oppression of women.

The dualistic concept of probability was close to the center of this epistemological transformation, being fundamentally involved in both the fracturing and the splintering of knowledge, as it were. At the same time that it helped to create the skeptical problem of induction, it provided the key to a solution, in the form of statistical inference. Through three centuries of fantastically rapid cultural change, these concepts have exhibited a hardness familiar only to theologians and

¹ Collingwood's (1956) belief that “the history of thought, and therefore all history, is the re-enactment of past thought in the historian's own mind” (p. 215) has made him most popular among historians as an object of ridicule, with the quirks of his thought being usually attributed to his progressive illness. I find, however, criticisms like Fischer's (1970) more delightful than convincing:

Collingwood's method calls to mind the New England cod fisherman in Kipling's *Captains Courageous*:

When Disko thought of cod he thought as a cod. . . [He] thought of recent weather, and gales, currents, food, supplies, and other domestic arrangements, from the point of view of a twenty pound cod; was in fact, for an hour, a cod himself, and looked remarkably like one. (Fischer, 1970, p. 196)

entomologists. To understand their extraordinary persistence requires an understanding, not only of the epistemological framework in which they were embedded, and which made them possible, but also of the wider cultural issues reinforcing that philosophical matrix. Among the latter, for example, was the sharp polarization of gender which emerged at about the same time. This polarity was aligned with many other dualities, and the position of statistical inference on the positively valued masculine pole helped greatly to make it as successful as the cockroach—or the Colorado beetle.

2.1 Brief Excursus on Historical Cognitive Change

The novelty of seventeenth-century developments has itself been the subject of some controversy; the shape of the debate and perhaps the underlying sentiment are reminiscent of certain other theoretical disputes—for example, that between Velikovsky and his critics, or between the nineteenth-century saltationists and evolutionary gradualists, or, perhaps, between the stage theorists of ontogenesis and the incremental learning theorists. Even though it should be clear that the concept of historical discontinuity is not to be taken in the literal, mathematical sense of yielding infinite second derivatives, the discontinuity hypothesis is nevertheless—whatever you get when you cross a red herring with a lightning rod—certain to derail at least some readers from the main argument. Consequently, I think even a very brief consideration of two previous such putative discontinuities will help to clarify what happened in the seventeenth century, not only in terms of form, but also in some ways, it turns out, in terms of content.

2.1.1 *Ancient Greece*

The first such cognitive revolution of which we have clear evidence began in ancient Greece around the time of Hesiod and the pre-Socratic philosophers and culminated in the work of Plato. Julian Jaynes (1976) traces the changes through the writings, perhaps spanning several centuries, which are commonly attributed to Homer. The developments in question can be summarized as a shift to interiority, from the concrete to the abstract, and from a participative mode of knowing to a sharp separation between knower and known. There is good reason to suppose that the introduction of writing was primarily responsible. Hellenic culture of about 1000 B.C. appears to have been unusual in being so highly developed while still lacking a system of writing. Greece may have been less rich in clay or papyrus than Sumer or Egypt, but the Cretans had a linear script as early as the seventeenth century B.C. (Diringer, 1968). A more likely reason for the “delay” may be just the success of the splendid alternative developed by the Greeks for the transmission of culture, namely, the oral tradition of epic poetry. The need for some such medium was heightened by the

scattering of the Greeks in the Dorian invasions around the end of the twelfth century B.C. Havelock (1963) shows in some detail how artfully embedded within the narratives were numerous observations on how things were done, from resolving conflicts to mustering a ship's crew, so that the epics constituted a sort of "tribal encyclopedia." Their poetic structure, moreover, was designed to facilitate memorization. Acoustic rhythms, for example, were reinforced in various ways during performance by bodily participation.

The date of the invention of the Greek alphabet has been the subject of considerable debate (cf. Diringer, 1968), but reliance on oral recitation of epic narratives persisted for several more centuries in any case. The Greek alphabet had the advantage of being as efficient as any system of writing that ever attained widespread use. It is also remarkable in having been taken over rather directly from the Phoenicians, as an aid to commerce, with the result that the names of the letters were meaningless in the Greek language. It is tempting to speculate that the evident arbitrariness of this connection between spoken and written language facilitated the abstraction of the self from the world. Donald (1991) suggests, however, that the uniqueness of the Greeks lay simply in their having been the first to record their mythic narratives, as opposed to using writing merely for shipping invoices, calendars, genealogical trees, and the like.

Writing brings its own limitations, which were disparaged at the time—as Havelock (1963) notes, "You cannot flourish a document to command a crowd" (p. 127)—but, as an alternative medium for the preservation of culture, it carried an invitation of its own for certain further developments. It imposes fewer restrictions on form; the permanence of the record does not depend on inciting a personal identification or emotional participation in the audience. More significantly, it affords us the opportunity to step back from what we have said and review it. The permanence of the record facilitates comparisons across instances, the isolation of common characters, reflection about abstractions which are not easily visualized, and the replacement of vivid verbs of action with sheer being, the reduction of predication to copulation. The abstraction of the self from the flux of experience proceeds correlative with the objectification of a stable world. Once the epics could be written down, the perceived inconsistencies of the narrative—for instance, the situationism of the implicit "ethics"—begged for resolution and systematization; and the intense appeal to the passions, now superfluous, came to be perceived as an actual distortion of the message. Thus Plato's call, on Havelock's account, was for detached, dispassionate reasoning, for reflective self-awareness, for philosophy in place of poetry. The object of knowledge was henceforth what was permanent, persisting through changing appearances. Seeking a name for the new, abstract object of knowledge, Plato chose one—*eidos*, form—which emphasized objectivity and externality, and effectively shifted the model of cognition from auditory to visual. The choice was not accidental. Havelock suggests ways in which Plato could have styled himself as "developer of the old" rather than as "prophet of the new" (p. 267), but Plato sought instead to separate himself as far as possible from the participative mode of knowing. Havelock wisely observes that "this is ever the way with makers of revolutions" (p. 266). The capacity for reflection and abstraction incontestably represented a

cognitive advance; the problem for Greek philosophy, at the time and in all its revivals, was just the vehemence of Plato's rejection of previous ways of knowing. Fortunately, but curiously, our practice has uniformly been better—less one-sided—than our preaching.²

2.1.2 *Medieval Europe*

Whenever thoughts of cognitive losses from impending senility threaten to depress us, we can take comfort, if we wish, from the Dark Ages, which remind us that history, even with its Arabs as repositories, is scarcely more secure than ontogenesis from the loss of precious developmental achievements. Radding (1985) discerns already in Augustine's thought a concreteness that contrasts with Cicero's, and he goes so far as to suggest that it was the cognitive deterioration of Roman civilization that made it more susceptible to invasions rather than vice versa. In the centuries that followed, the alphabet was not lost, but literacy certainly was, and along with it evidence of a capacity for abstract reasoning. By the sixth century, discourse had degenerated to parataxis; a narrative of a murder by Gregory of Tours—a light of his time—is overwhelmed and obscured by what we would regard as irrelevant details, omitting those we might think important, like the identity or motive of the murderer, all juxtaposed with no visible sense of organization (Radding, 1985).

The end of the early Middle Ages is usually dated from the rediscovery of the Greeks, though Radding thinks certain cognitive developments were necessary before the Greek authors would have made any sense. Writing and geographical dispersion again seem to have been involved in setting the stage for the cognitive revolution which occurred around the eleventh century. Bede, at the turn of the eighth century, had access to two different Latin translations of the Bible; being thus forced to grapple with the issue that the Scriptures were supposed to be God's exact words entailed some distancing from the text and an attitude of critical reflection. Similarly, the unification of all of central Europe in the Carolingian conquest a century later revealed an unsuspected diversity of religious belief and practice, making it difficult for people to continue in their assumption that they were all following God's laws. The embarrassing proliferation of conflicting authorities eventually made it necessary for writers to develop arguments in support of their views and to take account of the perspective of their audience or adversary. Dawson (1950) contends that the first such example arose in the eleventh century, from tensions inherent in the concept of the Holy Roman Empire, with the Church as political as well

²Or not so curiously: It would be more accurate to read “Plato” as metonymic for the whole sprawling tradition which he spawned. Like most innovators, he is often accused of the sins, and particularly the dogmatism, of his epigones; William Irwin Thompson (personal communication, February 4, 1993) observes that Havelock could equally be charged with what he himself criticizes in Plato, and McDonough (1991) reminds us that Plato strongly opposed the mechanical application of rules and subordinated the rule of law to the judgment of a sensitive statesman.

as religious leader. Pope Gregory VII led the reform effort, attacking the practice of simony as emblematic of the corrupting influence of the political world on the spiritual:³

For the first time in the history of the West an attempt was made to enlist public opinion on either side, and a war of treaties [*sic*] and pamphlets was carried on, in which the most fundamental questions concerning the relation of Church and state and the right of resistance of unjust authority were discussed exhaustively. (Dawson, 1950, p. 159)

Radding finds a general collapse of traditional authority around the year 1000 (though he does not suggest a connection with the failure of millennial prophecies), with the beginning of the twelfth century marking a new era in Western Europe comparable to the fifth century B.C. in Athens.

The emergence (or re-emergence) of interiority is one result: Appeal to authority entailed interpretation, which led to a focus on the intention of the author. In the new focus on intention, Radding, a Piagetian, sees a parallel with the development of moral judgment in children today. By around 1200, for example, intentions as well as consequences were coming to be seen as relevant to morality, and children and the insane were accordingly exempted from the penalties for murder. In corresponding changes in the political realm, the basis of social organization shifted from obedience to agreement and contract; a system of law and bureaucracy gradually replaced an agglomeration of decrees; and crimes came to be understood as offenses against the state rather than against individuals. (Radding, as Whiggish as Piaget in his historiography, does not question whether the last of these truly constituted progress, though Solomon, 1990, surely would.) In epistemology, the shift to human agency and authority appeared as the possibility for a more constructivist theory of knowledge. In the extreme nominalism of Roscelin, universals were no longer metaphysical, as they had been for Plato; for Roscelin, as for Boethius, universals were simply names, “mere puffs of air—the transitory sounds produced by speaking” (Radding, 1985, p. 209).⁴

If the incipient relativism and constructivism of such thinkers did not lead to a crisis of knowledge and self-consciousness well before the seventeenth century, the reason was evidently just the self-absorption and impoverished observational base of a tradition relying, in an era of hand-copied manuscripts, on texts as the ultimate reality. Windelband’s (1901/1958) plea for a more charitable appraisal is telling enough:

The circumstantial and yet for the most part sterile prolixity of the investigations, the schematic uniformity of the methods, the constant repetition and turning of the arguments, the lavish expenditure of acuteness upon artificial and sometimes absolutely silly questions, the

³The success of Gregory’s efforts can be judged from the complaint of Bernard of Clairvaux a century later that “the Church has become like a robber’s cave, full of the plunder of travellers” (Dawson, 1950, p. 246).

⁴Admittedly nominalism could be characterized as only weakly constructivist. If Roscelin’s philosophy sounds as though it might have appealed to J. B. Watson, Windelband (1901/1958) argues indeed that nominalism, however unimportant in its own time (even on Roscelin’s own pupil Abelard), led eventually to the associationism of the Enlightenment.

uninteresting witticisms of the schools,—all these are features which perhaps belong inevitably to the process of learning, appropriating, and practising, which mediæval philosophy sets forth, but they bring with them the consequence that in the study of this part of the history of philosophy the mass of the material, and the toil involved in its elaboration, stand in an unfavourable relation to the real results. So it has come about that just those investigators who have gone deeply, with industry and perseverance, into mediæval philosophy have often not refrained from a harsh expression of ill-humour as to the object of their research. (p. 268n)

(In style as well as content, one can easily imagine Sigmund Koch having found here a model for his commentaries on twentieth-century American psychology.)

The intellectual culture of the Middle Ages was not to be lost, as Hellenic-Roman culture was for 800 years, but would provide instead, in its own time, a point of violent departure for the next revolution. Deprived of empirical nourishment, the dry husks of scholasticism might well have simply blown away in the first breeze of the Renaissance; instead, the powerful support of the Church ensured a more schismogenic resolution. As Kuo Hsiang put it, in the third century, “A big thing necessarily comes about in a big situation” (Chan, 1963, p. 326; one assumes the Chinese expression was more elegant).

2.1.3 General Observations

It will be instructive to notice similarities and differences between these two early epistemological discontinuities—culminating in the fifth century B.C. and the twelfth century A.D., after two or three centuries of preparatory developments—before proceeding to that in the seventeenth century.

Either of these could be counted, for its particular culture, as the emergence of self-consciousness (or at least a step in that direction). In ancient Greece, the change was apparent even in language, as in the shift in loci of agency from *thumos* and *phrenes* to *psyche* (J. Jaynes, 1976). In the twelfth century, we see a new self-awareness reflected in the appearance, or reappearance, of phenomena as diverse as guilt and individual portraiture (Berman, 1989). Alternatively, and perhaps a shade more cynically, what I am calling the emergence of self-consciousness could be construed as the emergence of epistemology, an explicit concern with knowledge as no longer taken merely for granted. Both historical transformations also entailed a degree of trauma, inasmuch as the previous pragmatic-intuitive mode of cognition carried an unreflective certainty or assurance which was lost to the vagaries of philosophy. What was furthermore lost with the ascendance of philosophy was, in both cases, the legitimization of carnal knowledge and of participative, ecstatic experience, in particular.

The cardinal difference appears to be the context out of which they arose: The cultural tradition fought by Plato was well established and successful, whereas that confronted by Abelard or Anselm was by comparison mute and unformed. The consequences for subsequent development depend greatly on whether an existing order

must be dealt with. The typical, though not logically necessary, response in this case, as Havelock (1963) noted, is the most vehement possible opposition and denunciation. This was the pattern in ancient Greece and in seventeenth-century Europe, and it characteristically engenders deep splits and polarizations throughout the culture. Either mode of change, however, leaves its “shadow” side; Radding (1985) notes that as saints and devils gave way increasingly to natural explanation among the intellectuals of the twelfth century, mysticism and occultism found correspondingly renewed energy.

It should not be supposed, again, that any of these shifts was literally either instantaneous on a human time scale or total in its pervasion of the culture. Not only did the changes mobilize heretical resistance, but they also took time to trickle “down” from the intellectual elites who instigated them.⁵ But the gradual and partial character of any historical cognitive change should not obscure the qualitative differences that result.

It could be argued, following Foucault (1973b) and Radding (1985), that the turn of the nineteenth and twentieth centuries—roughly, the Kantian and Einsteinian revolutions—were quakes of comparable magnitude to those discussed here. It makes no difference to the argument of the present book whether they were or not. Only that in the seventeenth century is directly relevant to the development of statistical inference, but even the very cursory treatment here of the two preceding epistemological transformations may help to make this one less mysterious.

With regard to the seventeenth century, the discontinuity hypothesis would suggest either the occurrence of cataclysmic events in the preceding period or an extreme instability or ripeness for change in existing arrangements. Examples of both can be found. The fifteenth and sixteenth centuries certainly offer no shortage of events from which major sequelae would be expected: the discovery of the New World, the Black Plague, and the Reformation come immediately to mind. Other events of a much smaller scale may have been equally far-reaching: the invention of the printing press or the construction of visual perspective in painting, for example. Still more subtle effects may have been important. The early decades of the

⁵The vertical “*décalage*” resulting from such processes is visible in Western societies today. The collective sociability characteristic of the Middle Ages in general, for example, survived among the so-called working class until the twentieth century (Ariès, 1989), and the shared meanings and referents of a close social group make it possible to get along in everyday communication with what Bernstein (1971) calls restricted rather than elaborated code. Tough (1977) has thus found striking class differences in certain language features among children aged 3–7. When asked to describe a scene, socioeconomically disadvantaged children proceeded, like Gregory, by paratactic enumeration of concretes: “There’s a boy and a girl and there’s a bus, a red one, and some traffic lights.” Their advantaged counterparts not only used more abstractions but also projected more into the past or future and imputed a central meaning: “I think there’s going to be an accident because this boy is just going to run across the road when the bus is coming” (p. 99). The difference obviously has little directly to do with money and much to do with the extent to which reading and writing are part of everyday life. Print pulls for elaborated code, and for the first several centuries after its invention, literacy continued to be restricted to an elite. The democratization of education, particularly in the United States, had barely narrowed the class gap, however, when differential attachment to electronic media virtually reinstated it.

seventeenth century were a period of extreme cold, and Tyrone Cashman (personal communication, May 17, 1991) suggests, with some plausibility, that the change of climate may have contributed to the peculiar temper of the times. Merchant (1980) points, with a straight face, to a (related?) shortage of manure during the same period—leading to famine and social unrest. As examples of the instability of existing arrangements, the near-circularity of the Renaissance epistemology of resemblance (Foucault, 1973b), or the vulnerability of the fifteenth-century Church to *derrière-garde* attacks, could be cited.

In a purely structuralist approach, it might suffice to characterize the transformation in question merely as the substitution of one object of anxiety for another. Throughout the Middle Ages, Europeans were preoccupied with physical safety (Duby, 1988b). The vast forests between towns were like some neighborhoods of American cities today: It was commonly understood that to venture into them alone was to invite whatever misfortune occurred and to mark oneself as mad. And indeed, no sooner had the land been civilized and problems of physical security tamed than anxiety about the mind erupted everywhere. The simple substitution of objects recalls Foucault's (1973a) account of the emergence of madness, with the insane filling the social vacuum left by the disappearance of leprosy. Whatever merit such a global characterization may claim, for the purpose of this book, it will nevertheless be important at least to indicate the specific issues involved and their ultimate relevance to modern psychological research methodology.

A comprehensive account of the epistemological and wider cultural changes during this period would span a good many volumes. Trying to say it all in as many pages means that this chapter suffers simultaneously the defects of both extremes: being very concentrated while still very meager in its coverage of everything. In understanding such widespread cultural change as occurred in seventeenth-century Europe, the starting point also seems almost arbitrary: Whichever thread you pull on, the whole ball of twine goes into a knot. Foucault (1973b), in his analysis, took the transformation of the concept of sign as pivotal. To the degree, however, that human values are aptly synopsized by the popular slogan, “Money may not be everything, but it’s way ahead of whatever is in second place,” we might look to changes in the medieval economy as more fundamental.

2.2 The Concept of Skeuomorphosis

Money is a good example of what Theodore Porter (1992a, 1992b) calls technologies of distance, or of scale, and of a general process I see happening throughout history. Interaction at a distance pulls strongly toward the surface; the interior, the subtle, the idiosyncratic yield to the overt, the quantifiable, and the common as the basis for transactions. Since we can readily encounter things, or persons, only on the outside, we tend to lose sight of their interior or depth and come to take the shell for the reality. I am still looking for a good word for the process, though it may be too complex to be captured so succinctly; metaphorizing a term from anthropology, I

shall make do here with *skeuomorphosis*, a process of coming to take on the shape of the container. Before returning to money and its role in the seventeenth-century transformation of knowledge, I want to try to ground the concept with brief reference to a few historically earlier examples.

2.2.1 Early Technologies of Distance

One early example of a technology of distance is language. It is obvious that language enormously facilitates communication, making possible extensive and immediate interactions with complete strangers. At the same time, all of us find it difficult, at least at times, to put our experience into words, in a way that we feel does justice to the experience. The challenge of trying, in Alan Watts's phrase, to "eff the ineffable" is especially familiar in the case of affective or esthetic experience. But once we embody the experience in words, they become its new reality; they are what we remember, even as we know they are not exhaustive of the experience. We are so accustomed to this process that we hardly notice the price we pay in the impoverishment of our experience and our memories that results.

Two other devices, clothing and personal names, may have become important with the advent of civilization, around 10,000 years ago, when people settled down into cities. There were for the first time in human history more people in one place than could easily be kept track of, and Ableman (1985) believes that both clothing and personal names were introduced for the purpose of identification. The concept of modesty was introduced by religions thousands of years later; clothing, apart from warmth and protection from the elements, appears to have been introduced, probably along with tattooing, as a way of stylizing one's identity. The motivation for clothing may not have been purely cognitive, as Ableman implies; the familiarity and trust that would have been part of life in nomadic tribes of 50–150 people were lost with civilization (literally, living in cities), and it would not be surprising if a new layer of protection were felt to be needed in the company of thousands of strangers. Indeed, I strongly suspect that the direction of historical causation was the opposite of what we have been told: that, in history as well as ontogenesis, chronically keeping ourselves covered up led eventually to feelings of shame, which could then be exploited by religion (cf. Acree, 2005). Clothing would still count, however, as a technology of distance. And it is clear, in any case, to what degree we have come to invest our identity in our wardrobe and in our names—and also how superficial both these aspects of our identity are.

I say clothing "may have" become important with the advent of civilization because Knight (1991) believes it originated much earlier, in the Paleolithic era. He cites evidence by White of beads and other ornaments found from this time and reasonably infers the use of textiles and skins that would not have survived. Knight thinks the motive would have come from the time, well before the advent of language, when human head size drastically increased, necessitating prolonged dependency of children. Carrying and nursing young, women were impeded as hunters; to

assure themselves meat, they organized monthly sex strikes, broken when the men returned with meat. These entailed, among other consequences, synchronizing their periods. In hunter-gatherer societies around the world and throughout history, Knight notes, it has been taboo for a hunter to eat his own kill; he must bring it home for the women to cook and distribute. A system of such strong taboos needed reinforcement by a wide array of rituals and adornment, demarcating periods of shunning and inviting sex. Clothing, on this account, is not necessarily a technology of distance, enabling individuals to cope with large numbers, but a collective power tactic to enable the species to survive. But rituals, like the technologies of scale, are subject to decay from within, becoming shells of themselves, losing over time their connections to the meanings that generated them. We are left with the noxious “battle of the sexes” at a time when it is mostly women’s work to procure meat from the supermarket and when the division of labor in child rearing need no longer exert its toxic effect on us all (cf. *infra* on Dinnerstein).

2.2.2 The Emergence of the Market Economy

Certain economic changes during the medieval period laid the groundwork for the cognitive revolution that was to follow several centuries later. The effects can be characterized globally in terms of (a) the development of mathematical and mechanistic models, and of relativistic thinking, as Joel Kaye (1991/1992) has brilliantly shown, and (b) the peculiar reductionist consequences of technologies of scale.

Western Europe until around the eleventh century had largely a barter economy, with very little money in circulation (Kaye, 1991/1992, 1998). Barter is pretty much limited to face-to-face transactions, and it is difficult. If I wanted a car in a barter economy, I would have to go around to everyone who had the kind of car I wanted and hope somebody had a ton of data to be analyzed. As David Friedman (1996) points out, the reason finding a romantic partner is so difficult is that it operates on the barter system: You not only have to find somebody who has what you want, but that same person has to want what you have. Money is a technology of distance *extraordinaire*; it makes possible transactions at a distance both geographically and personally.

Market prices prescind from all qualities save one, which they measure with astonishing readiness and precision—all without the conspicuous exercise of individual judgment. The impersonality and automaticity of the process impart a kind of social objectivity, whose advantages are attended, as with most technologies, by liabilities. For this and other reasons, technologies of scale, once they are developed, tend to dominate. In the case of prices, their efficiency with respect to barter means greater wealth. But they enjoy also the advantage of concreteness and specificity: In somewhat the same way, perhaps, that experiences which can be put into words are more easily remembered, the precision of prices awards them a salience with respect to more vaguely defined qualities. The result, in any event, is a tendency for prices to become constitutive of value more generally. A familiar example

is the businessman encountered by St. Exupéry's (1943/1999) Little Prince. In describing a house, the Little Prince gets no reaction at all until he mentions how much it cost, whereupon the businessman agrees that it must be a very fine house indeed. As later centuries were eventually to appreciate, thanks especially to Marx, the market economy carries the risk that individuals will come to experience themselves and each other in the same reductionist terms as they are defined by the system. And the risk, while not unavoidable, has been significantly realized: The skeuomorphosis of humanity into commodity yielded the famous alienation of modernity.⁶

In the eleventh and twelfth centuries, the Italian city-states, less encumbered by feudalism than the rest of Europe, began trading with people of the eastern Mediterranean for luxury goods from eastern Asia, like silks and spices. These were soon much in demand in the rest of Europe, which expanded its production of goods like wool and furs for trade (Rothbard, 1979/2011). This tremendous expansion of commerce in Western Europe depended on a new technology: on a corresponding increase in monetization (Kaye, 1991/1992, 1998). England had 10 mints in 900, 70 in 1000 (Crosby, 1997), probably because the first veins of silver were found in Saxony in the 960s (Pye, 2015). It was the Frisians who reinvented useful money and taught the idea to the Franks under Charlemagne. Interestingly, money had been used prior to that time more for political than economic purposes: for taxes, gifts to the emperor, and the like (Pye, 2015).

2.2.3 Mathematical, Mechanistic, and Relativistic Thinking

One of the eventual consequences of the emergence of the market economy for the seventeenth-century crisis in knowledge was that everything was found to have a price—even salvation, through the sale of indulgences. Hadden (1994) suggests that the concept of general magnitude was abstracted from the exchange of dissimilar commodities, and Kaye (1991/1992) discerns here the origin of the idea that everything could be numerically measured. Aristotle's sharp distinction between quality and quantity was undermined, and debates about measurement at the Universities of Paris and Oxford adumbrated those six centuries hence—whether, for instance, it is possible to measure love, or anxiety, or intelligence. The idea that all qualities were susceptible to mensuration led in turn to a geometrization of the world (Kaye,

⁶The objectification continues into subcategories of humanity. Women today justifiably complain of being treated as sex objects, without commonly realizing that they treat men just as frequently as wallet objects: The first thing most women as well as men in this country say, when asked about their concept of masculinity, is “breadwinner” (Faludi, 1991). Some years ago, a friend persuaded me to take a look at personals ads in the Sunday paper. I laughed at what I took to be a typo in the first entry under “Women Seeking Men”: “must be over 65.” My amusement turned to something else when I saw how many said that.

1991/1992)—every quality represented by a continuum, every continuum by a line—and to the formulation of mathematical models.

The ordering of things under the new economic system had the effect, moreover, of undermining in several ways existing theories of natural order. Henry of Ghent was still arguing, in the thirteenth century, that a dying horse was worth more than a healthy ass, as Augustine had argued that a mouse, being alive, was more valuable than a pearl. But prices stubbornly declared otherwise. Money also provided an index of personal worth quite distinct from nobility and thus subverted aristocratic rankings. Eventually, in the conflict between economic and natural order, the latter yielded; the result was a mechanistic view of the universe, which no longer required a Mover to keep everything going (though Jean Buridan, with a bow to theology, acknowledged that God could have endowed the universe at Creation with forces which kept it going).

The discovery that such a complex system required no outside intervention had still further consequences. Aristotle had distinguished two kinds of justice in economic transactions. Distributive justice—virtually the opposite of what it is taken to mean today—required essentially that the transaction be free and voluntary. Rectificatory justice pertained to post facto adjustment of an unjust transaction by a third party, so as to bring the holdings into equality. In the hands of some medieval commentators, the latter, which they referred to as commutative justice, came to be considered on an equal footing, engendering debates about whether prices should be determined by the free market or by fiat. The latter alternative naturally had its appeal to prospective third parties, but in point of fact was found to be largely superfluous in practice. The observation that the free market tended to satisfy both partners to the exchange after all, Kaye (1991/1992) suggests, may have been an early source of the idea that *judgment could be replaced by calculation*.

Princes struggled nevertheless to make the economy behave as they wanted it to, creating a series of rapid devaluations and revaluations of the currency in the first centuries of monetization (51 in France between 1355 and 1360). Coping with such rapid fluctuations entailed an explicit relativism, thinking in different frames of reference (Kaye, 1991/1992). At the University of Paris, Buridan and Nicole Oresme were captivated by the implications of relativism wherever they turned their gaze. Oresme considered that there could be another universe at the center of our own (or, equivalently, that we could be inhabiting such a universe); Buridan recognized that, in the relative nature of motion, we could not tell whether it were the sun or the earth that moved. These two were far ahead of their time; the rest of the world would have to await more concrete and dramatic evidence, which was another century or two in coming.

2.2.4 *Scientific Objectivity*

The tremendous expansion of scientific work and exchange in the seventeenth century depended, as Porter ([1992a](#), [1992b](#)) has nicely analyzed, on a corresponding technology of distance, in this case the familiar concept of scientific objectivity. A technology for the conduct and coordination of science on a large scale will pull inexorably for the universals of experience, for abstraction, formalization, and impersonality. In a provocatively formulated but straightforward implication of constructivism, Daston ([1992](#)) observes that scientific communication was more nearly the precondition for the uniformity of nature than the reverse.

Quantification offers the most obvious instrumentality for scientific communication at a distance. Indeed, hopes for mathematics, as a universal language, went somewhat further: At least from the time of Leibniz and Hobbes, it seemed to hold the key to peace, in eliminating ambiguity and dispute. The press for the development of formal systems closely paralleled the quantification of the sciences. Deductive logic had been codified, however inelegantly, with the medieval taxonomy of the syllogism and related structures; the formalization of knowledge was to be completed with the formalization of inductive reasoning in statistical inference.

The technology of distance in science also entailed a standardization of observers, as uniform, interchangeable, impersonal, and, ultimately, invisible. This conception obviously paralleled the emerging political doctrines of egalitarianism and atomistic individualism, just as the older notion of quality work depending on skill and connoisseurship was coming to feel uncomfortably aristocratic. The conflict in ideologies may not have led to outright war in science, but neither was the transition abrupt or monolithic: Daston ([1992](#)) reminds us that reports of meteorite falls were discounted by the French Academy because they came from peasants.

None of these technologies of scale—quantification, formalization, observer exclusion—is intrinsically objectionable, any more than was the development of a commercial economy. They were subject, however, to the same skeuomorphosis as the price system. The alluring precision of numerical measurements gave them a saliency which tended to submerge the vaguer, qualitative aspects. Scientific work would ideally be reduced to meter reading. The emergent concept of objectivity depersonalized, decontextualized, and democratized knowledge, somewhat as money decontextualizes and universalizes value. One of the legacies of the seventeenth century is a banking metaphor for knowledge, which has now become nearly synonymous with data, bits to be stockpiled.

The desiccation of entities into shells of their former selves, into mere constellations of attributes, was to be completed by twentieth-century Anglo-American psychology, which reduced persons to points in space, the intersecting dimensions of variation skewering them through the breast like a butterfly on a card. Similarly, going beyond the seventeenth-century fascination with formal systems, the formalization of the sciences in the twentieth century emptied them of their content (O'Neill, [1991](#)). And we have moved beyond standardizing or depersonalizing observers to forcing their disappearance altogether. We write our scientific articles

in the third person to disguise our own contributions and make it appear as though everything just happened—though, Daston (1992) notes wryly, we still sign them.

These later developments will be discussed further in the last chapter. For now, we can observe the overall movement toward the surface in the changing conceptions of signs, causality, and representation. The combinatorial calculus of signs provides, moreover, a paradigmatic illustration of the hope for formal systems, and the incipient disembodiment of the observer will be apparent in the ascendance of the visual metaphor for knowledge.

2.3 Some Consequences for Epistemology

2.3.1 *The Concept of Sign*

It is widely supposed that the impact of the Renaissance was to shift the object of knowledge from the books of men to the book of nature. Garber and Zabell (1979), seeking continuity and stability, protest that the metaphor of the great book of nature had been around for centuries. The important point for the present context, however, is just that, whatever its remote precursors, the metaphor did not catch on until the Renaissance. The Protestant Reformation obviously helped to undermine ecclesiastical authority; it is also reasonable to suppose that the printing press, through its drastic cheapening of the written word, contributed to the break-up of the safe, ingrown scholastic fascination with texts. Certainly today, with millions of books representing different viewpoints on every conceivable subject, it is much more difficult for us to imagine that they all contain the sacred truth, than if all we possessed were a few ancient scrolls handed down for uncounted generations.^{7,8}

The epistemology of the Renaissance remained, like that of the Middle Ages, fundamentally a matter of interpretation, preserving the religious as well as the hermeneutic orientation. It was still God's word which was read, even if the text had changed; initially, at least, it was not religion, but only the clergy, which lost in the

⁷ Maclean (2002), citing work by Arrizabalaga, offers a piquant statistic on the abruptness of the change: The percentage of books published by live, as opposed to dead, authors rose from 0 in the 1470s to 40 in the 1490s.

Printing, however, like other technologies of scale, inevitably carried its liabilities, among them that *eo ipso* it made possible the concept of paperwork. Van Creveld (1999) says that, all over Europe from 1550 to 1650, monarchs had to create central state archives to keep up with the mountains of paper being generated by the growing bureaucracies.

⁸ The alert reader will have noticed, as mathematics textbook authors like to say, that the effects of the printing press were duplicated very closely, almost exactly 500 years later, by the emergence of the Internet. Overnight everyone who wanted to speak had a platform, and there was no overarching authority. Those who had presumed themselves the authorities were now challenged from all sides and reacted with a grace comparable to the Catholics of the Counter Reformation: by the ruthless suppression of dissent. On the social consequences of the collapse of authority attending the new medium, see Note 21.

transition. In the original meaning of the term, everything was *legend*: People read the stars for signs of the future, read the surface features of the earth for signs of water below, read the urine for signs of liver trouble. The basis for interpretation, for identifying and reading natural signs, was resemblance. Oswald Croll, a follower of Paracelsus, argued, for example, that the resemblance between a walnut and a human head must be a sign of the medicinal properties with which nature has endowed the nut. Thus the rind covering the bony shell ought to be useful in treating scalp wounds, and surely the nutmeat itself, “which is exactly like the brain in appearance,” would be good for internal head injuries (Foucault, 1973b, p. 27).

Foucault (1973b) speaks as though Renaissance epistemology were monolithic; but, although neo-Platonist hermeticism is distinctive of the Renaissance, it by no means replaced the medieval Aristotelian tradition, which, as Maclean (1998, 2002) insists, continued to flourish alongside it, especially in the universities. The distinction, like many others, can be overdrawn; the overlap is clear in the sixteenth-century university curriculum in medicine, which included semiology along with physiology, pathology, therapy, and hygiene (Maclean, 2002). Both epistemologies, in any case, had their vulnerabilities and wouldn’t survive the seventeenth century. The shift in the concept of sign, which Foucault sees as pivotal, actually served to undermine both.

Formally, signification was a ternary relationship comprising the significant, the signified, and their nexus. Since resemblances were both what signs might be recognized by and what they revealed, however, the three parts of the relation verged on collapsing into two, and the system was always perilously close to circularity. Ultimately the concept of sign itself was forced into a structural shift which, however subtle in appearance, was profound in its consequences. The change appeared in the fifth edition of the Port-Royal *Logic*⁹ (Arnauld, 1683), which introduced a new chapter on the concept of sign. Though Arnauld gave no acknowledgment that his formulation was new, it collapsed the middle term of signification into the significant, leaving a binary relation:

When one looks at a certain object only in so far as it represents another, the idea one has of it is the idea of a sign, and that first object is called a sign. Thus the sign encloses two ideas, one of the thing that represents, the other of the thing represented; and its nature consists of exciting the second by the first. (1683, p. 55)

“The signifying element,” writes Foucault (1973b):

has no content, no function, and no determination other than what it represents: it is entirely ordered upon and transparent to it. But this content is indicated only in a representation that posits itself as such, and that which is signified resides, without residuum and without opacity, within the representation of the sign. . . . Representation . . . is at the same time *indication* and *appearance*; a relation to an object and a manifestation of itself. (pp. 64–65)

⁹This book, *La Logique, ou l'Art de Penser*, was published anonymously in Paris in 1662. It is generally agreed to have been written by Antoine Arnauld, with some work in the early chapters contributed by Pierre Nicole. It went through many editions and translations and was used as a standard logic text at Cambridge, Oxford, and Edinburgh up to the nineteenth century (Buroker, 1993; Liddell, 1921/1973).

As a consequence of this simple shift, representation became coextensive with signification; all ideas were henceforth representations. Signification, now a relation between representations, rather than the voice of the world, became epistemological rather than metaphysical, as it were. Knowledge was effectively flattened into a single plane, with everything both signifying and signified. Hacking (1975) observes that, for Gassendi, even the middle term of a syllogism was a sign (though Gassendi was not the first; Emilio Campilongo made the same point explicitly in 1601; Maclean, 2002). Gassendi was not himself a skeptic, but it is apparent in his work what was happening to the scholastic idea of demonstration, when he could move the very model of proof to the realm of association and resemblance.

In the same movement, language itself became “dislodged” from things, as Foucault would say, and became an autonomous medium. In the Renaissance, words had been as natural as rocks and trees, on a footing with all other signs. Paracelsus knew, for example, that mercury was the proper treatment for syphilis, because it is the sign of the marketplace, where syphilis is caught (Hacking, 1975). In the centuries since, we have come to regard the connection between words and their referents as (mostly) arbitrary and would look upon the resemblance divined by Paracelsus as accidental. Already in 1662, in fact, the Port-Royal *Logic* was disparaging the notion of words as given, or natural: “The Ram, the Bull, the Goat could as well have been named Elephant, Crocodile, and Rhinoceros: animals which ruminate; therefore those who take medicine when the moon is in these signs are in danger of vomiting it back up” (Arnauld, 1662, p. 8). But in the time of Paracelsus, the whole problem was to find the correct names for things; then we should know what they were good for or what they meant. Words owed their usefulness to their fidelity to natural signs.

2.3.2 *The Combinatorics of Language and Thought*

The liberation of words from their bondage to resemblance meant that they now rooted their usefulness in a different ground, namely, in their exhibition of relationships to other signs (Foucault, 1973b). If language was no longer natural, however, neither was it arbitrary. The linguistic project of the seventeenth and eighteenth centuries was to construct a language, a system of representations, that would perfectly reflect the component elements of nature, so that the combination of signs in discourse might mirror the composition of nature. The ideal of language as a literal representation of nature in its component parts is reflected, for example, in the wish of Linnaeus that the text of his *Philosophie Botanique*, in its typographical modules, might exhibit a vegetable structure (Foucault, 1973b),¹⁰ or in the suggestion of Ehrenfried Tschirnhaus, in his *Medicina Mentis sive Artis Inveniendi Praecepta*

¹⁰Linnaeus may have taken his inspiration from the common practice of printing epicedia in shapes such as a heart (Sylla, 2006).

Generalia of 1687, that the true definition of laughter ought to be funny (Windelband, 1901/1958). (I don't know what he thought about the definition of error.)

The ideal of a linguistic calculus exemplifying a mathematical model of knowledge spanned the whole space of epistemologies in the period from 1650 to 1800. The appeal to Leibniz and Spinoza is obvious, but for Hobbes and for Locke as well, thinking consisted merely of combinations of verbal signs. This program, making logic coincident with semiotics (Windelband, 1901/1958, p. 452), is generally considered to have reached its ultimate expression in Condillac's *Logic* (1780/1822).¹¹

"The widespread interest in combinatorial problems from the early part of the seventeenth century," according to Kassler (1986):

coincides with the revival of materialistic atomism and the appearance of the mechanical philosophy of René Descartes. According to that philosophy, material elements in motion merely change their position or arrangement in time, thereby generating new configurations of material elements. (p. 39)

Windelband (1901/1958) traces this interest in combinatorics beyond Descartes to the Renaissance revival of neo-Pythagorean number mysticism, in which numbers must represent the ultimate reality. The original source, however, appears to have been the work of the thirteenth-century Catalan theologian Ramón Llull (see Erdmann, 1893). Llull's various combinatorial inventions included, for example, a series of concentric circles which could be rotated to bring into apposition different sets of words, represented by letters inscribed around the edge of each circle. In the simplest such device, 16 attributes of God could be paired, leading to the revelations that God's greatness is good (BC), God's goodness is great (CB), and so on. Llull wrote as though his contrivances could be used to answer virtually any question that could be raised.

It is commonly assumed that a major impetus for the development of Llull's lexical algebra was his effort to convert Moslems and Jews, for whom appeals to human authority were useless. Erdmann (1893) and Bonner (1985) note the delicacy of his task, since proofs or mathematical demonstrations have the effect of undermining faith; but by the seventeenth century Leibniz could make explicit the hope that calculation could replace argument and that mathematics could hence be used to compel assent. As Albury (1986) and others have noticed, a certain totalitarian ambition thus lurked in the mathematical program.

Descartes, Bacon, and their contemporaries ridiculed Llull, but they had read him, and Bonner (1985) discerns in their work more of a resemblance than they acknowledged. Swift is thought to have been satirizing Llull in his Laputan professor who invented a device with words printed on different faces of cubes, various combinations of which were exposed by turning a crank: "The most ignorant person at a reasonable charge, and with a little bodily labor, may write books in philosophy,

¹¹ Windelband's odd remark that Condillac's book was designated as "a textbook for 'Polish professors'" (p. 478n) recalls the equally curious title of McCall's *Polish Logic* (1967). Windelband was merely making cryptic reference, however, to the fact that Condillac's work was commissioned by the Polish government for the reform of its educational system.

poetry, politics, law, mathematics, and theology, without the least assistance from genius or study” (quoted in M. Gardner, 1958, p. 2). The pervasion of Enlightenment thought by combinatorics can perhaps best be seen by comparing the report by Kassler (1986) that there was published around 1763 “A tabular System whereby the Art of composing Minuets is made so easy that any Person, without the least Knowledge of Musick, may compose ten Thousand, all different, and in the most pleasing Manner” (p. 33).

In science, the method of analysis and recombination according to genetic principles found its most natural application in the chemistry of Lavoisier, who modeled his writing explicitly on the work of Condillac. In other fields, the application was more strained, though eighteenth-century botanists had a serious go at it. Adanson, in his *Familles des Plantes* in 1763, suggested that botany was closely tied to mathematics, inasmuch as plants are distinguished “only by their relations of quantity, whether numerical (discrete) or continuous, which tell us the extent of their surface or their size, their shape, their solidity” (p. cc; my translation from Sheynin, 1980, p. 326). He, and Auguste de Candolle after him, envisioned a botanical algebra which would make it possible to calculate, say, the species that lay exactly halfway between a cucumber and a geranium (Foucault, 1973b; Sheynin, 1980). Such efforts may strike us nowadays either as risibly quaint or as impressively modern, depending on our appraisal of present-day multivariate statistical analysis.

In fact, if the idea of reducing thought to calculation has a familiar ring, it is because it has been picked up by the program of artificial intelligence. Windelband could not easily have imagined Condillac’s philosophy surviving unchanged through the Kantian revolution into Anglo-American philosophy a century hence, and being taken as a model, once again, for human cognition. The project of artificial intelligence has foundered on the problem of semantics (Dreyfus, 1979; H. Gardner, 1985; Rychlak, 1991), which is just the old problem of resemblance. Instead of essaying a constructivist account of resemblance, in the manner, for example, of Werner and Kaplan (1963), it has attempted simply to replace semantics with syntactics, as the Enlightenment philosophers attempted to replace resemblance with the combinatorics of signs. Both systems assume the symbols given, and thus neither can account for its origins. (On the problem of origins, see below, pp. 57–60.)

2.3.3 *Causality*

In the new language of calculation, the relations among signs had become wholly external. As resemblance was replaced by relations of order and measurement, and interpretation gave way to representation, knowledge, now a matter of mere reckoning with signs, was losing its depth. The world was becoming a black box. This general externalization of knowledge would not have been possible, however, without the dissolution of the concept of cause, with which it went hand in hand. Formally, the concept of causality underwent a transformation in some respects like

that of signification: Both metaphysically and epistemologically, the two terms of a formerly asymmetrical relationship were flattened into the same plane, becoming formally similar.

Metaphysically, causality had been, in the Aristotelian conception, a relation between an entity and its actions. The mathematical approach of Kepler and Galileo, however, led to a shift from the study of *things* to the study of their *motions*. Hence, while even Descartes still held to the Aristotelian notion of *aitia* as substances or things, Galileo went back to the earlier Greek thinkers, like Leucippus, for whom causality was a relation between states, or motions. From Galileo on, “Causes are motions, and effects are motions. The relation of *impact* and *counter-impact*, of the *passing over of motion from one corpuscle to another*, is the original fundamental form of the causal relation” (Windelband, 1901/1958, p. 410).

Epistemologically, causality was assimilated to the omnivorous concept of sign: Causes were signs; effects were signs. Their frequency of co-occurrence was left as the only observable relation between things. Neither medieval meanings nor scholastic causes any longer held things together. Characteristics became detachable from entities and from each other; their association became a matter of statistical coincidence.¹²

The identification of cause and effect with signs may not have been explicit until Berkeley in 1710, but the elusiveness of causality to an epistemology of representation was already being felt in scientific work in the seventeenth century. Windelband notes indeed that the concept of causality had greater inertia in metaphysics than in science, which could more easily start afresh. The work of Francis Bacon was a watershed, spanning aspects of both sixteenth- and seventeenth-century epistemologies. He was influential in shifting attention from documents to empirical observations as the supporting base of knowledge, and his canons of empirical method fit perfectly the new epistemology of causality. What we know better through Mill (1846) as the methods of agreement, difference, and concomitant variations, Bacon described as the construction of *tabulae præsentiae*, *tabulae absentiae*, and *tabulae graduum*. Comparison of these instances of presence and absence would then lead, by a process of elimination, to the isolation of causes. The apex of Bacon’s pyramid of knowledge, however, was still the classical demonstration of causality; and Windelband (p. 384) characterizes Bacon’s own scientific work, as a search for formal causes, as entirely scholastic in its conception.¹³ His empirical program was the inspiration for the founding of the Royal Society of London in 1660; his

¹²A striking duplication of Humean skepticism regarding causality and inference is found in the *Tattvopaplavasimha* of Jayarasi Bhatta in the seventh century (though nonsurviving manuscripts on the Carvaka doctrine go back to the sixth century B.C.). As seventeenth-century empiricism came about in reaction to the inscrutability of agential scholastic causes, so the skeptical Carvaka philosophy of Jayarasi was directed against the extreme focus on spirituality in the Hindu orthodoxy. In Indian philosophy, however, an intense spirituality has perfused all other systems, even the heterodox, so that the Carvaka school has remained an isolated minority. See Radhakrishnan and Moore (1957).

¹³He apparently overlooked Bacon’s (1627) one excursion into experimentation; see Chap. 5.

successors in the Society, applying his methods, found that their results led them, almost in spite of themselves, to a more Galilean conception of causality. Instead of leading up Bacon's pyramid to the demonstration of formal causes, the celebrated advances of seventeenth-century science—for example, Boyle's gas laws—defined regularities and constant conjunction, rather than things (Hacking, 1975). Indeed, the natural philosophers of the Royal Society from the start shunned general inquiries into causality as “metaphysical”—a term that was already becoming opprobrious, as witness its use in criticism of the differential calculus (Jeffreys, 1961, p. 405n) or inquiry into the chemical composition of stars (Kneale, 1949).

Descartes was, like Bacon, a transitional figure. In his later writings, he came gradually to abandon the goals of certainty and the identification of causes:

Just as the same craftsman could make two clocks which tell the time equally well and look completely alike from the outside but have completely different assemblies of wheels inside, so the supreme craftsman of the real world could have produced all that we see in several different ways. I am very happy to admit this, and I shall think I have achieved enough provided only that what I have written is such as to correspond accurately with all the phenomena of nature. This will indeed be sufficient for application in ordinary life, since medicine and mechanics, and all the other arts which can be fully developed with the help of physics, are directed simply towards applying certain observable bodies to each other in such a way that certain observable effects are produced as a result of natural causes. And by imagining what the various causes are, and considering their results, we shall achieve our aim irrespective of whether these imagined causes are true or false, since the result is taken to be no different, as far as the observable effects are concerned. (*Principles of Philosophy*, cited in Franklin, 2001, p. 220)

Here, Franklin says, Descartes “asserts that there is no need to find the true underlying model to make a correct prediction. It is the first clear statement of the dream of modern statistical inference: to make true predictions independently of difficult inquiry into inner causes” (p. 221). He adds: “The modern economic modeling that attempts to forecast unemployment, interest rates, and so on without any commitment to grand economic theories is a continuation of Descartes’ project” (p. 221).

The disappearance of causality was thus partly involuntary, as it were, but also, almost as if to make a virtue of necessity, partly voluntary. It happened as an unavoidable consequence of the epistemology of representation and the philosophy of mechanism; but, if causes were inaccessible to the Baconian, or Millian, canons of experimental inquiry, then they were necessarily unscientific and to be eliminated where they intruded. And intrude they did, just because formal philosophy was at this point pulling away sharply from everyday experience and its articulation. The concept of causality was not rejected as meaningless; rather, it was banished to the same metaphysical limbo as similitude. Neither could be denied outright. Descartes himself recognized that any comparison for purposes of ordering depends on the apprehension of some dimension of similarity. Similarly, the compilation of instances to derive or to support empirical laws is parasitic on some concept of causality beyond sheer juxtaposition in time and space. The problem was essentially that the epistemology of representation left no accredited methods for discerning either resemblance or causality. Because the concepts were not repudiated outright, but remained meaningful on some level, modern epistemology was left with a residual consciousness of its superficiality. The primary goal of experimental inquiry is

still supposed to be the identification of causal relations; the fact that that task is impossible has led to systematic equivocation.

It has been a major task of philosophy in the last two centuries to polish that equivocation to the point where it became invisible. In the eighteenth-century calculus of representations, it was virtually entailed that all predication be reducible to an equals sign; and, indeed, we find very much this sort of statement, reminiscent of Thales, in Destutt de Tracy's *Éléments d'Idéologie*, published at the turn of the nineteenth century:

The verb *to be* is found in all propositions, because we cannot say that a thing *is* in such and such a way without at the same time saying that it is. . . . But this word *is* which is in all propositions is always a part of the attribute [predicate] in those propositions, it is always the beginning and the basis of the attribute, it is the general and common attribute. (cited in Foucault, 1973b, p. 96)

Twentieth-century analytic philosophy would add mainly the terminology of set theory and class inclusion. The meaning of cause and effect is virtually taken to be exhausted by protasis and apodosis, and the concept of implication itself has been replaced by the splendidly named concept of material implication. The whole point of the latter concept, of course, is that it is anything but material, and causality can now be represented in a truth table, in a simple juxtaposition of presence and absence, which can in turn be represented electronically by arrays of bistable multivibrators.¹⁴

On the other hand, in the older, pre-Enlightenment view, even propositions making the most nakedly reduced of attributions still carried connotations of causal connection. The Greeks, for example—for whom truth, beauty, and love were more of a piece—evidently did not mean, in asserting “All *A* is *P*,” merely to be reporting the results of an empirical survey, real or hypothetical; rather, they sought to link the attribute *P* with the nature of *A*, to assert the existence of an intimate connection between *A*-ness and *P*-ness.¹⁵ Something like this has remained closer to the everyday meaning.

In a certain sense, as Windelband noted, science could still go on about its business without being held up over semantics. Yet, over the long run, philosophy does to some extent dictate the terms to science; and, while scientific practice, in a formal sense, has not changed much over the last 300 years, philosophy has moved to a position very much farther from everyday discourse, from which it issues its challenge. Science, in the middle with respect to language, is clumsier in its equivocations on causality, but remains committed to them. Thus talk of causality in science is still shunned as metaphysical and naïve, but the stigma attaches, curiously, only to the first term of the relation. It is still entirely standard to speak of experimental *effects*, as in “Effects of Stress on Susceptibility of Albino Rats to Sound-Induced Convulsions” (Griffiths, 1962). A similar asymmetry characterizes an ancient pair of euphemisms, *antecedent* and *consequent*. The former, devoid of causal

¹⁴ Some more recent work on causality in artificial intelligence is discussed in Chap. 10.

¹⁵ Cf. the formulation of the Aristotelian logician H. W. B. Joseph (1916):

An universal judgement connects a certain attribute with a certain subject in virtue of their nature and without regard to the frequency of their existence. . . . The causal relation which connects *a* with *x* connects a cause of the *nature a* with an effect of the *nature x*. (pp. 408–409)

connotations, implies that temporal succession is all that causality amounts to; its logical complement, *subsequent*, is unsatisfactory because it gives the whole duplicity away. The success of *functional relationship* is of course due to its perfect ambiguity: *Function* retains two separately respectable meanings, one purely mathematical, the other irrefragably causal.

2.3.4 *Representation*

As scientific knowledge, in its formal description, was pushed ever farther from the meanings of everyday discourse, it tended to become *sui generis*. This divergence was part of a more general linguistic movement in which, according to Foucault (1973b), history began to pull apart from science, and literature emerged as a new form. In the sixteenth century, when words were as much a part of the world as herbs, comets, or stones, he says:

To write the history [in the sense of the word at that time] of a plant or an animal was as much a matter of describing its elements or organs as of describing the resemblances that could be found in it, the virtues that it was thought to possess, the legends and stories with which it had been involved, its place in heraldry, the medicaments that were concocted from its substance, the foods it provided, what the ancients recorded of it, and what travellers might have said of it. (p. 129)

From the mid-seventeenth century on, the documents and fables were no longer part of the thing itself, and the web of meanings connecting it to the world; they became its history, in the modern sense.

Meanwhile, as scientific discourse was becoming more restricted, literature arose as a kind of counter discourse—not to speak of different content, but to absorb the expressive, nonrepresentational functions of language. Our culture still acknowledges, if only partially, that literature, which deals frankly in metaphor and word play, interpretation and resemblance, has something to tell us about the nature of things, some deeper meanings to reveal; but, out of a deep distrust of resemblance, it has tried to keep literary methods strictly out of science. Metaphor, as Mailer (1966) says, is confined to the “ghetto of poets” (p. 308).

The scientifically knowable world, the glory of the Enlightenment, had become a world of arbitrarily associated surface features, without organic connection. By the beginning of the nineteenth century, particular sciences would discover the principles of organic structure and function; Foucault (1973b) indeed dates the emergence of modern biology, as well as economics and linguistics, from this point. It is no accident, he would suggest, that botany should have been paradigmatic of eighteenth-century science. Construed as essentially a taxonomic enterprise, it lent itself with particular plausibility to the analytic epistemology of that period. Physiology, on the other hand, could become active only in the nineteenth century. But whereas the organic revolution, as it were, spread throughout the Western scientific community, in philosophy it was largely restricted to the Continent and to Germany particularly. (The evident implication that philosophy is tied more closely

than science to culture is another challenge, if one were needed, to its traditional assumption of hegemony over the sciences.) In Anglo-American philosophy, as was noted above, the appeal of the mechanical model, of parsing and recombining a given set of representations, was so strong that it would still constitute the information-processing mainstream of epistemology in the twenty-first century.

The plausibility, and longevity, of the representational theory had much to do with vision, with its peculiar properties, having been taken as paradigmatic of knowing. The visual model goes so deep that it is very hard for us today to “envision” alternatives, but it was neither an inevitable nor an inconsequential choice. Keller and Grontkowski (1983), in a paper aptly titled “The Mind’s Eye,” claim, interestingly, that the verbal form of *know*—*ksuniemi*—used by Heraclitus originally meant to perceive, especially by hearing.¹⁶ The visual metaphor, as they point out and was noted above, began with Plato. In his theory, the eye participated in vision by sending out its own stream to the object, and vision arose from the harmony of streams emanating from the eye, the object, and the sun. In an era so passionately committed to empirical observation and distrustful of the occult (though not, of course, without a notable revival of interest in the “shadow,” such as was long an embarrassment to Newton’s biographers), it was perhaps inevitable that vision should be seized on again as the model of knowledge. But in Descartes, particularly, the participative, communing aspect of vision—as experienced, for example, in the phenomenon of “locking eyes”—was lost, and sight became a wholly passive process. Descartes probably did more than anyone else to undermine the emanative theory of vision and to promote the copy theory of perception, and this latter theory accorded very well with the emerging representational theory of knowledge.

The problem for the passive copy or representational theory is the perceptual or cognitive processing that occurs on the way to awareness. It is the peculiarity of vision that it could have permitted the illusion in the first place of “unprocessed” knowledge, of a literal copy or reflection. By the same token the visual model also lent plausibility to the distinction between primary and secondary qualities—and to

¹⁶They refer for support to von Fritz (1974), where I find, however, no mention of *ksuniemi*. The literal meaning of the word is to bring or set together (Liddell & Scott, 1843/1968); its metaphorical meaning is often translated as *comprehend*. The negative form *axynetoi* occurs in the first and second fragments, where Heraclitus complains of not being understood: e.g., “Although this account holds forever, men ever fail to comprehend, both before hearing it and once they have heard” (Kahn, 1979, p. 29). But the closest von Fritz comes to discussing *ksuniemi* is in relation to Fragment B114, in which Heraclitus makes a pun on *xyn nooi* and *xynoi*: “Those who speak with *noos* [*xyn nooi*] must base [what they say] upon that which is common [*xynoi*; cf. *koinon*] to all and everything” (von Fritz, 1974, p. 38). Heraclitus argues that few people have *noos*, by which he means insight into the divine law governing everything, and furthermore that most people do not understand the truth when they *hear* it. But *xynnoia*—meditation—is a different word; most tellingly, von Fritz contends: “It is hardly without significance that there is no fragment which admits as much as that anyone ever has or could have grasped the truth merely by hearing it” (p. 40). And despite the obvious fit of the temporal extension of the sense of hearing with the supposed process philosophy of Heraclitus, Kahn contends that Heraclitus’ favored word for *know* was *ginoskein* (whence *gnosis* and *cognition*).

the very concept of primary qualities, as those which were intrinsic to the object, and in no way dependent on the form of the consciousness experiencing it. The troublesome implications then lay with secondary qualities, those which did depend on consciousness. On a superficial level, their refractoriness to quantification—it is easier to say how long something is than how cherry it tastes—posed a major challenge to the Cartesian epistemology of order and measurement; and their subjectivity (though it was not yet called that) renewed vigorous efforts at quantification, which have scarcely abated since (Rüstow, 1980). On a deeper and more important level, the contrast with primary qualities implied that consciousness is inherently distorting. Paradoxically, then, the very activity of consciousness became the disqualifying element of knowledge.¹⁷ So it is that the representational theory tends inevitably to harden into what Neil Friedman (1967) calls the “dogma of immaculate perception” (p. 111).

The illusion of noninvolvement of the body in awareness,¹⁸ as well as the more general fallacy of observer exclusion, can be attributed to two or three special properties of the visual modality. First, vision, like hearing, is an allocentric sense (Schachtel, 1959), which operates at a distance; but with vision, unlike hearing, distance is often regarded as an asset. We speak of stepping back for a wider perspective; and the greater our remove, the wider our field of vision. Second, in contrast to hearing, vision presents the world outside of time and thus lends a kind of support to the notion of eternal truths. As a third such feature might be counted the common assumption, which appears to be shared by Keller and Grontkowski themselves, that an object is unaffected by our seeing it—though Arthur Koestler (1972), referring to the demand for actors who can keep one on camera, would suggest penile erection as a counterexample. The more obvious “processing” involved in the sense of taste, on the other hand, would have sufficed to spoil its appeal as a model for knowing.

The visual model then—and a particular “view” of vision, at that—gives a unique plausibility and concrete reality to a conception of knowledge as disengaged, detached, objective, universal, and timeless. Other sensory modalities have been used, minimally, as models for knowledge; we mean something important when we speak of being “in touch with” some aspect of reality. But if the more body-bound, autocentric senses had been taken as synecdochic, the project of grounding cognition in biology (e.g., Maturana & Varela, 1980) and in the body (e.g., M. Johnson, 1991; K. J. Shapiro, 1985) would not have been so very late in coming, and the dream of artificial intelligence would not have seemed so compelling as it has (e.g., Dreyfus, 1979; Winograd & Flores, 1986).

¹⁷ Gebser (1953/1985) points out that the Latin *mens* (mind) and *mentiri* (to lie) come from the same root.

¹⁸ Etymology reminds us, if we have forgotten, of the somatic origins of our concepts of knowing and being: *Be* comes from the Sanskrit word for breath; *nous* originally pertained to the sense of smell; *sapere* referred to flavor; and *inspiration* still refers both to breath and thought (cf. Onians, 1951).

2.4 Some Psychological and Cultural Considerations

2.4.1 The Emergence of Self-Consciousness

The exuberance of the escape from the monasteries to the real world is reflected in the explosion of discoveries and inventions from the Renaissance onward, which were to have profound consequences, in their own right, for the concept of knowledge. These included a dramatic change in the dimensions of the known universe and of the position of the human species within it.

The discovery and exploration of the New World, for one, confronted Europeans with a multitude of cultures, both advanced and primitive by their standards, with different languages and systems of thought. The existence of diverse languages and cultures had always been known; but, so long as it was only a few strange neighbors one was dealing with, it would have been easier to maintain the belief that one's own language and culture were the correct ones. And indeed initial reports of different cultures on newly discovered continents were evidently greeted at first much as they might be by modern tourists, as something like museum specimens, with no particular implications for one's own world view. The earlier mode of encounter was purely assimilative: When Charlemagne introduced the Gospels to the Vikings, he made Jesus a Northerner rather than a Jew—one of those swarthy Southerners who were already held in low repute; that is presumably the source of those ubiquitous dime-store portraits of Jesus, 1000 years later, as brown-haired and blue-eyed. By the time of Montaigne, however, recognition of different systems of thought had become explicit (Bordo, 1982/1983). Arriving just as signs were withdrawing from things into the space of representation, this dislodging of ethnocentrism cannot have but aided the emergence of language as epistemological or of knowledge as relative to the knower.

Any lingering ethnocentrism or anthropocentrism was further assaulted by the heliocentric theory of Copernicus (no longer the sheerly fanciful suggestion of Buridan) and by the invention of the microscope and telescope, which revealed unsuspected microcosmic and macrocosmic worlds without apparent limit (Kearns, 1979). The well-known existential terror of Pascal before the “jaws of infinity” contrasts startlingly with the adventurous attitude of Oresme:

When I consider the brief span of my life absorbed into the eternity which comes before and after—as the remembrance of a guest that tarrieth but a day—the small space that I occupy and which I see swallowed up in the infinite immensity of spaces of which I know nothing and which know nothing of me, I am terrified and am amazed to see myself here rather than there; there is no reason for me to be here rather than there, now rather than then. (quoted in Cashman, 1974, p. 76)

Bordo (1987) suggests that the real impact of the Copernican revolution was the shift from a centered to an acentric universe, so that, on a psychological level, the eventual result of all these dislocations was a loss of a sense of place, of belongingness, in the world.

This felt consequence was reinforced, she further suggests, by a subtler and more profound development: the fifteenth-century discovery of visual perspective in painting.¹⁹ Like monetization (which may have helped to bring it about), visual perspective contributed to a geometrization of space, which served thereafter to situate objects relative to one another, but no longer so much to contain them. A world held together by meanings, in a vast net of resemblances, gave way to the world as an array of objects. The newly abstract sense of locatedness generalized beyond the spatial to the temporal: It was in the sixteenth century that books regularly began to show the author, date, and place of publication (Ariès, 1960), as if it were felt that these things now needed to be marked. Thus it was, perhaps, that the concrete dislocation produced by geographical exploration and technological advances was undergirded in theoretical terms by the newly abstract sense of space and time.

In focusing attention in another way on locatedness in space, recognition of visual perspective, again like monetization, facilitated thinking in relative terms: the awareness that perspective changes from one location to another, hence that different perceivers are given different worlds. The impact of this awareness was heightened by the circumstance of vision having been taken as a metaphor for knowledge more generally. Thought, no less than sight, must be relative to vantage point. Indeed, “perspective” has become for us virtually a frozen metaphor for a theoretical framework.

It is not a long step from appreciation of the relativity of perspective to the doctrine of the fallibility of sense perception. Inasmuch as signs, language, and knowledge in general had lost their former connection to the world, it is not surprising that error should have become epistemological also. Bordo (1982/1983, 1987) points out that, for the ancient Greeks, error or illusion was attributed to discrepancies in the outside world—for example, between the world of particulars and the world of forms—rather than to defects of consciousness as such. With the incipient doctrine of primary and secondary qualities in Galileo and Descartes, later formalized and named by Locke, distortion became intrinsic to consciousness. Hence, just as attention was focused on the epistemological finitude and solitariness of individuals,

¹⁹ Pierre Francastel (1951/1977) incessantly reminds us of the ethnocentrism entailed in referring to the new spatial conception of the Italian Renaissance as a discovery. The realist view of linear perspective is as absurd, he suggests, as the notion that a single language might suffice for all the semantic needs of humanity. Space is constructed, as Pinol-Douriez (1975) and others have shown, and cultures differ in their constructions. The particular geometric view of the Renaissance Francastel attributes to the supersession of the view of the world as a concrete representation of God’s thought by the notion of it as a reality in itself, with its own eternal attributes. Kubovy (1986) offers a different cultural explanation for the emergence of perspective, namely, that Renaissance painters, in creating a center of projection and forcing it to diverge from the viewer’s standpoint, separated the mind’s eye from the bodily eye, as it were, and induced thereby a spiritual experience that could not be achieved by any other means. Whatever the specific explanation, it is interesting to note, with Francastel, that the anachronic notion of the abrupt discovery of eternal truths in the Renaissance was first articulated by Vasari in the sixteenth century, just at the time of the Counter Reformation and of Roman imperialism.

their cognitive apparatus was found to be defective. Small wonder, then, that the Cartesian search for indubitability should have been so pressing.

These various developments taken together—the shift in the status of language from metaphysical to epistemological, from given to constructed, and the recognition that what is seen depends on the observer—could be said in a serious sense to constitute the birth of subjectivity, the emergence of self-consciousness for the species. This would make it the third time in Europe alone, on the present account; but in each case, as was noted above, the historical transition, though identifiable as a distinctly new phase, remained yet incomplete; and the second was, in any case, but a partial recovery of what had been lost from the first. As an explicit philosophical construct, in fact, self-consciousness would await the Kantian revolution, and its implications are far from fully realized even now. In the seventeenth and eighteenth centuries, self-consciousness appears more as an implicit theme, coming closest to articulation in the anxiety of Pascal.²⁰ The emergence of self-consciousness for the species was presumably attended by no less a sense of unease than the corresponding experience on the individual level, in ontogenesis; but several additional circumstances made it unfortunately much more threatening than it might otherwise have been. These had especially to do with the effect of the Protestant Reformation on the epistemological authority of the individual. The epistemological impact of the Reformation was complex and far-reaching, and its effects were both beneficial and destructive.

2.4.2 Polarizations, Alignments, and Projections

In the popular view, the meaning as well as the instigation of the Reformation is represented in Luther's heroic opposition to the sale of indulgences. In Rüstow's (1980) view, on the other hand, Luther's challenge actually constituted an attack from the rear, as it were, on a Church that had become liberalized in some positive

²⁰At this level, the phenomenon was not limited to intellectuals; manifestations can be discerned in everyday life as well. The obvious example is mirrors. Mirrors of mica, bronze, or glass were used in antiquity, as long as 6000 years ago by the Egyptians; they were popular in Greece and Rome, then virtually disappeared until (no surprise) the twelfth century (Berman, 1989). The century then saw an explosion of mirror production and significant technological improvement. Developments in etiquette and privacy could be adduced as further examples. Manuals of etiquette appeared at this time, as did individual eating utensils and dining chairs in place of benches. Outdoor activities moved indoors, and interior space became more differentiated (Contamine, 1988). The construction of hallways, beginning in the fifteenth century, obviated passing through one room to get to another (Ariès, 1989). Beds, which had measured up to 11 feet (de la Roncière, 1988), no longer had to sleep the entire family. Public nakedness and sex, which had been tolerated at least in special circumstances like baths, became unacceptable, even as objects of conversation. The social construction of self-awareness generalized to institutions serving to articulate group differences. Seventeenth-century efforts toward standardization of grammar and orthography (Caput, 1972), for example, give the impression of having been motivated in part by the desire to mark social class distinctions.

ways and not merely politically corrupted. As a specifically religious force, the Church had declined up to the Renaissance, and its gradual melding with pagan religions paralleled the formation of new languages, like French, from Latin and indigenous roots. Without the Reformation, the Church might ultimately have lost its independent identity, as Renaissance humanism set about to unify and harmonize science and religion. As it was, the Church was forced, with the Counter Reformation, back into a conservative position more extreme than at any point in its history. Christianity, in contrast to pagan religions, had always been strongly directed to the afterlife and scornful of earthly concerns and pleasures; but, under the ascetic influence of Calvin in particular, that orientation was intensified. The result was alienation not only from the earth but from self: “Calvin expressly preached the religious duty of self-hatred—‘haine de nous mesmes’” (Rüstow, 1980, p. 283).

Another of the roots of the Reformation, as well as its important legacies, was the notion, propagated by fourteenth-century dissident theologians John Wycliffe and Jan Huss, that individuals were competent to interpret the Bible for themselves. This notion, rather opposite in its thrust, was on the whole, of course, a tremendously beneficial, liberating move, for it extended the epistemological authority of the individual by implication to the pursuit of knowledge generally. With authority, however, comes responsibility. The burden is a serious one for anybody, at any time; but at this point in history, it was being shouldered just as sense perception was being found deficient and the base of knowledge in general was being uprooted, and as self-reliance was being damned and self-abnegation exalted.

Inevitably, and unfortunately, the new epistemological license also instigated a virtual dogfight. Within the Protestant domain, nobody any longer had any special authorization, no way to adjudicate conflicting claims; and, given the supposed stakes of eternal salvation or damnation, everybody was maximally threatened and insecure. The well-known results included an efflorescence of religious factions, engaged in perpetual war and persecution over petty doctrinal disputes, as well as the formation of defensive alliances. Rüstow (1980) remarks dryly that:

It would be possible to write a history of scientific and intellectual progress during these centuries without mentioning the Reformation and Counter Reformation. Only in the biographical notes on the bearers of this development would either movement make itself occasionally embarrassingly noticeable in the form of emigrations, persecutions, trials, jailings, tortures, and executions. (p. 301)

Certainly the pervasive atmosphere of fear and oppression cannot be disregarded in understanding the intellectual productions of the time. As Karl Pearson (1978) notes, “Anyone who pronounced the word ‘tolerance’ was considered an atheist” (p. 439); and Rüstow adds that atheism, for Calvin, was a capital crime: In the first 5 years of his rule in Geneva, he had 10 persons beheaded and 13 hanged. His most famous victim, Michael Servetus, who anticipated by nearly a century William

Harvey's discovery of the circulation of the blood, was burned for the crime of antitrinitarianism.²¹

A remarkably large proportion of the philosophical writing of the seventeenth and eighteenth centuries was devoted to providing a rational basis for religious belief. Even the first significance test (Arbuthnott, 1710) was an argument for the existence of God. Out of context, this large literature would appear curious, since there were precious few voices defending atheism. Endemic persecution is of course part of the explanation, but Hunter (1981) adds that *atheism* was used very broadly at that time to cover heterodoxy of any kind, including simple deism, in addition to loose morals. In the 1620s, the Pilgrims of the Plymouth Colony denounced as atheistic "the use of the Anglican Book of Common Prayer, the maypole, and selling rum and firearms to the Indians" (Rothbard, 1979/2011, p. 154). The new science was particularly subject to atheist accusations, as natural explanations threatened to leave no room for supernatural ones. The Cambridge Platonist Ralph Cudworth found the rejection of final causes atheistic (Hunter, 1981). A very few thinkers, like Hobbes, dared to carry the implications of the new natural philosophy to its evident conclusions, but most, like Boyle, were defensively concerned to show it as revealing God's handiwork.

The impressive advances of science, on the one hand, and the rift and Counter Reformation within the Church, on the other, led to some interesting realignments. Both seventeenth-century science and religion followed the model of the medieval Catholic Church, of forming protective alliances with the state, but all three institutions jealously guarded their own domains. The new science was threatening to the state as well as to the Church: for, although the Baconian program had been the inspiration for the establishment of the Royal Society, it also included an activist utopian agenda for social reform. The solution achieved was for the state to charter and support the Society (albeit minimally, in a financial sense, for years) while also carefully circumscribing its domain of activity. Van den Daele (1977) quotes Robert Hooke's draft Statutes of the Royal Society: "The Business and Design of the Royal Society is: To improve the knowledge of natural things, and all useful Arts, Manufactures, Mechanics, Practices, Engynes and Inventions by Experiments (not meddling with Divinity, Metaphysics, Moralls, Politicks, Grammar, Rhetoric, or Logick)" (p. 31). The search for a safe, neutral ground for science ended in its being defined in terms of method. This three-way compromise between church, state, and science thus provided for official sponsorship of science by the state, with the

²¹The authority of the Church having declined dramatically from the sixteenth century to the twenty-first, the role of the sixteenth-century Church was by that time taken over by the secular Academy, which spoke with a remarkably uniform voice, on matters of ideology as well as science. Heterodox views are dismissed as "conspiracy theories"; they are not supported by federal grants; they have trouble getting published; they are blocked from social media platforms. The general atmosphere of intimidation is powerful enough that the few dissidents to speak up are typically retired or at least tenured. As of early 2021, it appears that the forces of orthodoxy, proclaiming themselves the true champions of democracy, will press on toward total annihilation of dissent. The culture seems more radically split than at any point in its history, and there will be no scientific revolution to help us.

church and state jointly invested in controlling its activities. In the three centuries since, of course, the role of the church has declined, while state sponsorship and control of science have become very much greater; but the success of that compromise is attested by the fact that method is to this day taken as definitive of science.

Its success has in fact obscured who the losers were in that arrangement. But there was one form of inquiry which disappeared at that time from accepted science: the “natural magic” of healers and thinkers like Paracelsus. Not only did their methods not fit with the new science; but, because of claims that Christ might have been a natural magician, Protestants and Catholics alike perceived such work as an atheistic threat; and Paracelsus’ sympathy with the peasants also made him a political menace. Thus church, state, and science joined hands to defeat Renaissance humanism. For a century or so thereafter, whatever was magical was necessarily diabolical and to be destroyed. In the new orthodoxy, supported by both religion and science, matter had no special, secret powers, and the earlier, magical mode of communion with things was lost.

It may seem strange that science and religion should have needed each other as well as the state. Certainly Luther’s contempt for the “harlot reason” left little opportunity for a merging of the two, and they were instead thrust sharply apart. But, thus alienated, each was weaker on its own. As they carved up their separate realms, Easlea (1981) suggests, their common rejection of natural magic helped to save both science and Christianity, in mutually estranged forms.

It would be fair, in fact, to characterize the seventeenth-century changes in knowledge in general as a process of polarization. The entire field was suddenly riven by dualities: science vs. religion, mind vs. body, self vs. world, subject vs. object. Reuniting these fragments was to be a major project for philosophy for the next several centuries. We are still so thoroughly caught in those splits that it may be difficult to conceive how radical and alien they were for their time; but it is worth pondering how such changes, precisely in the direction of *disintegration*, could have taken over with such force, how they could so quickly come to seem natural and to be taken for granted.

The general dynamics of revolutions, noted above, may be mainly responsible, but in one respect, I believe, the influence of religion, and the Reformation in particular, has been largely overlooked. That is the profound *moral* rejection of the body by Christianity—its antipathy to the body, to sexuality, to materiality in general, to life on earth as opposed to the hereafter. The position of the Church had been clear for over a thousand years; no less a figure than Augustine had condemned the pursuit of knowledge as “the lust of the eyes” (Pusey, 1966, p. 229). But it had also become more liberalized and benign over that span, until the time of Calvin, when asceticism and mortification of the flesh again prevailed. Rüstow (1980) notes that Calvinism destroyed “merry old England” once and for all; the casualness and comfort with the body in the time of Rabelais were not to be known again in Europe. I think it is not unreasonable to assume more than a coincidental connection between the epistemological rejection of the body in the seventeenth century and its recently renewed moral rejection. It would necessarily have appeared to Descartes that to be *res cogitans*—an entity of pure consciousness, without material extension or

contamination—was the only way to be moral. Descartes, Hobbes, Leibniz, and Spinoza were not conventionally religious for their time; but neither, apparently, at least for most of them, was their professed deism merely a ploy to avoid persecution. Perhaps it was, in any event, that doctrines such as the dubitability of sense perception felt right on a level deeper than epistemology.

Rejection of the feminine It would be astonishing, in an era so pervaded by value-laden splits, if the corresponding attributes were not projected onto actual groups of people. The principal losers in this case were women and, less obviously, children. The nascent polarizations with respect to gender and age underlay and reinforced all the others, especially the splits between mind and body and between public and private.²²

If men experienced themselves as composed, divisibly, of soul and body, they tended to project each of these, whole, onto women. The Judeo-Christian tradition is unusual among religions in lacking goddesses, even if it is not with respect to the distinctly secondary status accorded to women. Christianity also features a sharp split in its image of woman: Mary and Eve, the virgin and the temptress. This image appears to have been enhanced in the sixteenth and seventeenth centuries. In terms of social roles, there had evidently been a rough equality between men and women in the Middle Ages, at least for all but the aristocracy (Bordo, 1982/1983, pp. 262–263); in fact, there was little distinction between men's and women's clothing (Braunstein, 1988).

Toward the end of the fifteenth century, however, the image of women, and of nature, shifted from just and nurturing to unpredictable, secretive, and withholding. Of the various explanations that have been offered for the suddenly extreme persecution of women, one of the more plausible, and less noticed, is the severe epidemic of syphilis that lasted from about 1500 till the late seventeenth century (Andreski, 1982, 1989). There is good reason to believe that the disease was brought back by Columbus' sailors; in any event, a newly virulent form appeared in Europe at that time. The variable course of the disease and its symptoms conducted to occult interpretations, and the psychosis of its advanced stages simulated bewitchment. The specifically psychological threat was to the clergy. Although chastity had been increasingly encouraged through the Middle Ages, it was not enforced until the Council of Trent (1545–1563) outlawed the common practice of concubinage. Syphilis removed the deniability of transgression; the concept of witchcraft restored it.

Explanations of witch burning merely as a device for the subjugation of women must account for the fact that 10–20% of the victims were male, including young

²²With respect to gender issues in this period, it is crucial to bear in mind that what was written was overwhelmingly the work of men. Merchant (1980) notes that, because custom excluded women as authors, when Anne Conway's *The Principles of the Most Ancient and Modern Philosophy* was published after her death by her friend and editor Francis van Helmont in 1690, the title page listed van Helmont as the author. Conway's book, which was the inspiration for Leibniz' monadology, was still erroneously attributed to van Helmont by scholars 200 years later.

boys. For that purpose, moreover, burning is overkill; Andreski points to Moslem culture as evidence that the humiliation of women can be accomplished by more subtle means. The syphilis hypothesis also accords with the specific duration of the witch burning craze. However benighted and superstitious the preceding centuries, witch burning had not been a significant part of the culture; Charlemagne, denouncing it as a pagan custom, had prescribed the death penalty for it. The twin epidemics of disease and burning also disappeared at about the same time.

Mind and body The persecution of women in the sixteenth and seventeenth centuries, whatever its source, has always been well known; what remained curiously invisible until it was uncovered by recent feminist scholarship was the relation of that persecution to science. The close identification of women with the body, with matter, and with nature, however, implicated them inexorably in the new epistemology. The new science, and mathematics especially, required the greatest possible remove from the passions—from the body, in general, which was the source of error. Joseph Glanvill (1665/1885), a Fellow of the Royal Society and a propagandist for it, wrote as a rather glib and caustic skeptic: “For us to talk of *Knowledge*, from those few indistinct representations, which are made to our grosser faculties, is a *flatulent vanity*” (p. 168). But he made it clear that it was the female element that was to blame:

Where *Will* or *Passion* hath the casting voyce, the case of *Truth* is desperate. . . . The *Woman* in us, still prosecutes a deceit, like that begun in the *Garden*: and our *Understandings* are wedded to an *Eve*, as fatal as the *Mother* of our *miseries*. (1665/1885, p. 99)

We love the issues of our *Brains*, no less than those of our *bodies*: and fondness for our own *begotten notions*, though *illegitimate*, obligeth us to maintain them. We hugge intellectual deformities, if they bear our Names; . . . and if we might determine it, our proper conjectures would be all noted *Axioms*. Thus then the *Female* rules, and our *Affections* wear the breeches. (p. 114)

Women, it seemed obvious, were unfit for the profession of science. Condillac, in the mid-eighteenth century, believed that girls literally had softer brains and were therefore more impressionable (Daston, 1988). They were also effectively barred from practicing medicine, through a device that was to remain a model for related professions, including psychology, to the present: A classical education was made a legal prerequisite for practice. There is reason to wonder whether Greek and Latin, or even the science of the day, made men much more effective healers,²³ but the requirement was effective in excluding women as competitors.

The political persecution of women was paralleled by the subjugation of the body and the metaphorical rape of nature in science. For the Renaissance alchemists, the root metaphor of knowledge was coition: the union of mind and matter, of female and male principles. Paracelsus, for example, saw medicine as the wife desired by the disease (Keller, 1985). By the time of Bacon, the metaphor had

²³The necessity of doctors’ knowing Greek was actually questioned as early as the 1530s (Maclean, 2002).

switched to rape. In Bacon's case, this result could be explained, at least in part, as the inevitable consequence of combining the existing view of nature as feminine with a vigorous experimental activism. But the fervency of his rejection of "mere experience," and of his demand for controlled conditions of observation, was perhaps a little extreme: "There remains mere experience, which when it offers itself is called chance; when it is sought after, experiment. But this kind of experience is nothing but a loose faggot, and mere groping in the dark" (Montagu, 1850, p. 357), which (he returns to the metaphor later) "rather astonishes than instructs" (p. 363). It was on this basis that he scorned the scholastic disputes, which "catch and grasp at nature, but never seize or detain her: and we may well apply to nature that which has been said of opportunity or fortune, 'that she wears a lock in front, but is bald behind'" (p. 368):

We intend not to form a history of nature at liberty and in her usual course, when she proceeds willingly and acts of her own accord, (as for instance the history of the heavenly bodies, meteors, the earth and sea, minerals, plants, animals) but much rather a history of nature constrained and perplexed, as she is seen when thrust down from her proper rank and harassed and modeled by the art and contrivance of man. (p. 341)

It is necessary, however, to penetrate the more secret and remote parts of nature.
(pp. 345–346)

If any individual desire and is anxious . . . to *know* a certainty and demonstration, let him, as a true son of science, (if such be his wish) join with us; that when he has left the ante-chambers of nature trodden by the multitude, an entrance at last may be discovered to her inner apartments. (p. 344)

(According to Franklin, 2001, in the 1590s, Bacon was involved in investigating Catholic plots and so would have witnessed torture.) Collingwood (1956) adds his own specifically sadistic gloss:

The scientist must take the initiative, deciding for himself what he wants to know and formulating this in his own mind in the shape of a question; and . . . he must find means of compelling nature to answer, devising tortures under which she can no longer hold her tongue. Here . . . Bacon laid down once for all the true theory of experimental science . . . [and] of historical method. (p. 269)

The rape metaphor was scarcely confined to Bacon; as both Easlea (1981) and Keller (1985) amply show, it pervaded seventeenth-century writing on the new science. Nor was Bacon the most extreme; Keller is careful to note that, for all his graphic sexual imagery, Bacon still occupied an intermediate position of some complexity: for "Nature," after all, "is only to be commanded by obeying her" (Montagu, 1850, p. 370). But even in Bishop Sprat we find only a slight softening of the language of Machiavelli (cf. Merchant, 1980, p. 130):

The man chosen to defend the Royal Society against its critics and detractors, Thomas Sprat, spoke (as beffited a future bishop) not of the rape of nature but rather of a forceful wooing: "Whoever will make a right, and a fortunate Courtship to *Nature*, he cannot enterprise, or attempt too much: for *she* (as it is said of other *Mistresses*) is also a Mistress, that soonest yields to the *forward*, and the *Bold*." And what would be the successful outcome of such bold courtship? "The Beautiful Bosom of *Nature* will be Expos'd to our

view,” Sprat rejoiced, “we shall enter into its *Garden*, and taste of its Fruits, and satisfy our selves with its *plenty*.” (Easlea, 1981, p. 85)

The prevalence of such rape fantasies might be explained in part by the more general need for control, subjugation, and humiliation of the body. But it should also prompt a search for cultural sources of sexual humiliation of men. Possibly the reports coming back from African expeditions about races with exceptionally large penises had their effect in undermining men’s sexual security. Keller (1985) notes that, in general, gender distinctions were enhanced and that women were specifically forbidden to imitate men in their hairstyle, dress, or behavior. (See also Lucas, 1988, on female transvestism during that period.)

The more important influences, however, were surely ideological and moral. Certainly the doctrines of Luther and Calvin inculcated powerfully a sense of sexual shame, which could only have increased tensions between the sexes. Easlea (1981) notes Thomas Browne’s regret that there was no way of reproducing like trees. Male sexuality more specifically came to be the target of ridicule. The codpiece, which had come into fashion in the fifteenth century, disappeared during the seventeenth. The sexual act itself came to be seen as ridiculous and beneath the dignity of a gentleman, as indeed perhaps it is: The concepts of ladies and gentlemen seem specifically to exclude whatever is sexual about women and men, but the former defined the new standards of conduct. A certain view of women (for instance, in the *Malleus Maleficarum*) regarded them, it is true, as so insatiable that they had to resort to copulation with the devil; but, even here, the sexual asceticism and frustration to which Calvinist teachings would be expected to lead may provide the explanation: for it is not hard to imagine in men—especially errant clergy exposed by syphilis—anger and a need to be controlled that were disowned and projected onto women.

Easlea (1981) proposes outright that science constituted a surrogate sexual activity, more becoming a gentleman of class. Certainly there were plenty of references to the new philosophy as “masculine.” The interesting paradox noted by Keller (1985) could be taken in support of Easlea’s hypothesis: that scientists were simultaneously perceived as asexual and hypermasculine, as if the latter made up for the former. On the one hand, science was (and still is) regarded as a masculine activity in comparison with, say, artistic professions. Esthetic pursuits require being intimately in touch with the affects and passions and the feminine side of things, from which science ideally carries us as far as possible. Not only are scientists themselves supposed to be less sexual and passionate than artists; there may even be some basis for such a perception of them. Karl Pearson (1978) observes that there is almost a rule that great mathematicians never marry.²⁴ In the seventeenth and eighteenth centuries, it would cover Pascal, Descartes, Newton, Leibniz, de Moivre, and d’Alembert. Lagrange did take what his friend d’Alembert referred to as the

²⁴The reason may have been identified as early as 1700 by Bernardino Ramazzini, in his work on occupational diseases: “Almost every mathematician . . . is unworldly, lethargic, suffers from drowsiness, and is utterly impractical. The organs of mathematicians and their whole bodies inevitably become numb” (quoted in Sheynin, 1982, p. 247).

“perilous leap,” but neglected to mention it in his letters to d’Alembert because it was so unimportant (K. Pearson, 1978). Isaac Todhunter, several orders of magnitude less bright than these stars, also married, but, all the same, Kendall (1963) reports, he took Hamilton’s *Quaternions* with him on his honeymoon.

Public and private If the implications for knowledge of the gender alignment of the mind-body split were the more dramatic, the consequences of the split between public and private were no less important. On Jessica Benjamin’s (1988) argument that the public sphere is deeply identified as masculine and the private as feminine, we should expect to find the public-private distinction growing sharp at the same time as the other fissures probed in this chapter, and indeed that prediction is readily enough confirmed. Braunstein (1988) observes, for example, that writers of the fourteenth and fifteenth centuries often had trouble maintaining a distinction between public and private; a supposedly private memoir would be swamped with details of civic life. Duby (1988a) and Contamine (1988) implicate government in the centrifugation of these two spheres. In the feudal revolution around 1000, as strong central government collapsed, medieval society became, in modern terms, at once public and private throughout: Every household was, as it were, its own state. With the return of powerful and increasingly interventionist governments toward the end of the Middle Ages, however, the home became a refuge against the intrusion of the state. Various indications of a growing sense of privacy were mentioned above (Note 20) in connection with the emergence of self-consciousness.

The implications for knowledge of the differentiation between public and private derive from the specific nature of the respective poles. Whereas the private sphere is densely constituted of particulars, contexts, and dependencies, Benjamin (1988) characterizes the public realm as a masculine world of abstractions, of interchangeable, autonomous individuals whose behavior is governed by universal laws. In its political incarnation as the Enlightenment ideology of rights, this masculine conception of the public has enjoyed a remarkable success in view of the evident implication that blacks, women, and children, inherently lacking the unalienable rights of white men, were not human beings. The analogous public conception of knowledge psychologists will have no trouble recognizing as their context-stripping methodology, as Mishler (1979) calls it—still as vigorous and fresh as if it had been flash frozen three centuries ago and protected from any intervening intellectual growth. Even for the “other,” correlational discipline of scientific psychology (Cronbach, 1957), interest in individuals lies only in their position relative to others in the distribution. The basis for a more “feminine” model of knowledge already exists, Flax (1990) believes, in psychoanalysis, particularly in the analytic relationship; E. F. Keller’s (1983) illustration in the work of Barbara McClintock has become the standard (but see the critique by Richards & Schuster, 1989).

The rejection of roots As the spatial body was to be transcended by reason in the epistemology of the Enlightenment, so, too, was the temporal past. Like the physical body, the history of both the individual and the culture remained an anchor in the murky obscurity of material reality, an unwelcome reminder of finitude. The rejec-

tion of the past, however, was not only a denial of finitude, of situatedness in a particular context; it was also a denial of the incompetence and vulnerability of infancy. The latter, of course, was connected to the repudiation of the feminine.

In a century newly sensitive to the finitude of perspective, to the fact of situatedness in the world, in a particular body, with a particular history, denial held the appeal of an ersatz transcendence. Real transcendence, to the extent that it can be achieved, entails precisely the fullest awareness of self, of context. Whatever we learn about ourselves tells us that much about the world, about our construction of it: We know more what it is we see when we see where we stand. That awareness is essentially the project of psychoanalysis. If we deny our past, on the other hand, our blindness to it persists as a constant but invisible feature determining our horizon. Jane Flax (1983) eloquently describes the role of the latter dynamic in the philosophy of Descartes:

The posture of Descartes' cogito replicates that of a child under two in its relations to a caretaker. . . . One reaction and defense to the discovery of separateness is narcissism, in which the outside world is seen purely as a creation of and an object for the self.

Through "good enough" social relations this stage is transformed into a genuine reciprocity in which separateness and mutuality (interdependence) exist simultaneously. However, denial of separateness, of the individual integrity of the object (mother) will lead to the adoption of narcissism as a permanent character structure, precisely the type of solipsistic isolated self with delusions of omnipotence which Descartes' cogito displays.

Furthermore, underlying the narcissistic position, the fear and wish for regression to the helpless infantile state remains. The longings for symbiosis with the mother are not resolved. Therefore, one's own wishes, body, women and anything like them (nature) must be partially objectified, depersonalized and rigidly separated from the core self in order to be controlled. Once this position is established, the relationship between the self (subject) and object (other persons, nature, the body) becomes extremely problematic, perhaps unsolvable. This frozen posture is one of the social roots of the subject-object dichotomy and its persistence within modern philosophy. It is an abstract expression of a deeply felt dilemma in psychological development under patriarchy and thus cannot be resolved by philosophy alone. (pp. 260–261)

Bordo's (1982/1983) study of Descartes indicates that he was rather explicit himself in his wish to escape both history and childhood. In the first instance, he described himself as "a mind so far withdrawn from corporeal things that it does not even know that anyone has existed before it" (quoted in Bordo, 1982/1983, p. 195). And on childhood:

Since we have all been children before being men and since we have necessarily been governed for a long time by our appetites and by our teachers . . . it is almost impossible that our judgments should be as pure and solid as they would have been if we had had complete use of our reason since birth and had never been guided except by it. (p. 238, ellipsis hers)

For Descartes, she writes, "The state of childhood must be *revoked* through a deliberate undoing of all the prejudices acquired in it—and a beginning anew with reason as one's only parent. This is precisely what the *Meditations* attempt to do" (p. 239). She goes on:

Psychoanalytic theory urges us to examine that which we actively repudiate for the shadow of that whose loss we mourn. The culture in question had already *lost* a world in which the

human being could feel nourished by the sense of oneness, of continuity between all things. In the Cartesian separation of *res extensa* and *res cogitans* we find a decisive expression of denial of any basis for such a sense of continuity. (pp. 244–245)

The rejection of the past was not merely a Cartesian phenomenon. Joseph Glanvill was as eloquent on the subject as ever:

Our *Reasons* being inoculated on *Sense*, will retain a relish of the stock they grew on: And if we would endeavour after an unmixed Knowledge; we must *unlive* our former *lives*, and (inventing the practice of *Penelope*) undo in the *day* of our more advanc'd understandings, what we had spun in the *night* of our *Infant-ignorance*. (1665/1885, p. 63)

In fact, the program of obliterating the past and starting with a clean slate, like the combinatorics of the Idéologues or the empiricists, was a universal theme at the time, not limited merely to science and philosophy. Its consequences in the political realm were illustrated with particular vividness by the French Revolution.

Since the rejection of roots seems only slightly less subtle a theme in seventeenth- and eighteenth-century thought than the repudiation of the feminine, it is interesting that Foucault (1973b) should appear to have reached the opposite conclusion, that that period was characterized by a preoccupation with origins and with genesis. It was indeed at that time, he notes, that history came into its own as a discipline distinct from science. I believe the inconsistency in both instances is only apparent—that my claims about the rejection of roots can be squared both with what Foucault sees as a search for origins and with the emergence of history at this time—but it will be illuminating nevertheless to retrace his argument regarding the genesis of knowledge and to compare it with the present one.

Foucault's argument derives from the relation between resemblance and imagination in the thought of that time. It is imagination, he observes, that makes resemblance possible: “If representation did not possess the obscure power of making a past impression present once more, then no impression would ever appear as either similar to or dissimilar from a previous one” (p. 69). But resemblance is only the starting point of knowledge; it has to be analyzed into identities and differences and ordered comparisons. That imagination allows us to discover only resemblances, the source of confusion and error, may be attributed either to defects of the imagination, which would otherwise grasp identities and differences directly, or “to the disorder of nature due to its own history, to its catastrophes, or perhaps merely its jumbled plurality, which is no longer capable of providing representation with anything but things that resemble one another” (p. 70). Either attribution, Foucault argues, entails a search for a state prior to disorder and confusion:

The possibility of a science of empirical orders requires . . . an analysis of knowledge—an analysis that must show how the hidden (and as it were confused) continuity of being can be reconstituted by means of the temporal connection provided by discontinuous representations. Hence the necessity, constantly manifested throughout the Classical age, of questioning the origin of knowledge. (p. 73)

Curiously, however, Foucault's own examples are just as striking for their agenetic character: “the mythical form of the first man (Rousseau), or that of the awakening consciousness (Condillac), or that of the stranger thrust suddenly into the world (Hume)” (p. 70). These represent attempts, to be sure, to trace the

development of knowledge from its origin; but it is also notably an artificial, stipulated origin, one of which we are completely in control. It is a fresh start, without the impenetrable past that each of us, individually and culturally, actually starts with. It was natural for Richard Price to pick up Hume's hypothetical stranger thrust suddenly into the world when he introduced Bayesian statistical inference (Bayes, 1763; see Chap. 4), but the state of total ignorance that he presupposed is hardly relevant to human inference. To the extent, then, that genesis was a preoccupation of Classical thought, it was wholly in this abstract, hypothetical sense—a reconstructed genesis to replace the real one. And in that sense Foucault is right.

The emergence of history as distinct from science might, again, be thought to reflect a new concern for origins, and ultimately it would indeed make possible a focus on history as an object in its own right. Historiography of science, in any serious sense, is really, however, an achievement only of the twentieth century. Todhunter's *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace* (1865) offers one illustration of nineteenth-century history of science. It is essentially a paragraph-by-paragraph guide to works of that period, and accordingly, as Kendall (1963) observes, is "just about as dull as any book on probability could be" (p. 204). The natural philosophers of the seventeenth and eighteenth centuries paid their perfunctory respects to the ancient writers, but really saw themselves as empiricists starting out with a fresh view unencumbered by the doctrines of the past. Their attitude toward the history of science prevails 300 years later.

In a certain subtle but important sense, the separation of history from science could be regarded, not as an obvious advance in the growth of knowledge, but an actual obstacle to it. It was only in the postpositivist era of the late twentieth century that attempts were begun to understand science in its full historical context (see, e.g., the anthologies by Buss, 1979; Graham, Lepenies, & Weingart, 1983; and Schuster & Yeo, 1986). What is entailed by reconnecting science with its secular origins, however, is relinquishing our implicit image of it as a transcendental realm set above human foibles, which has suited it so well in the last century or so for its role as an ersatz religion.

It might be expected that the philosophical disowning of the past would have been paralleled socially by a renewed oppression of children. If there is no clear evidence for such a shift, it is only because, as de Mause (1974/1988) eloquently argues, children could hardly have been treated more inhumanely than they had been up to that point. What does appear to have happened enabled a more subtle form of oppression: the construction of a discontinuity between childhood and adulthood. As is well known, Ariès (1960) has argued that, whereas children previously had evidently been regarded very much as small adults, from this point on the attainment of both the age of reason and of sexuality marked a distinct threshold. Children were set apart, with a status less than human and in some ways more like that of pets: to be owned, controlled, trained, protected.

The oppression of women was not so different in form: Their exclusion from more and more kinds of work, and the concomitant glorification of home and motherhood, removed them conceptually to a safely distant domain. Women, being more powerful than children, had to be more actively oppressed, but the treatment of the

two was similar in the sentimentalizing—and therefore belittling—of their newly defined roles and a patronizing attitude toward them. The subtlety of the oppression may have made it more palatable to both sides; in any event, the general cultural acceptance of these roles made formal political oppression almost superfluous. In the United States, widely regarded until recently as a liberal society, a Constitutional amendment recognizing equal rights for women would be highly controversial 300 years later, and the proposition that children are human beings with the same political rights as adults (see H. Cohen, 1980, and Farson, 1974) would still be regarded as radical, if not preposterous. So it was, perhaps, that the oppression of children, in its peculiarly modern form, emerged in parallel with that of women, and was similarly linked to science.

2.5 Summary and Preview

The present chapter has ranged widely over sixteenth- and seventeenth-century thought and culture, by way of sketching the context and conditions for the emergence of the modern concepts of probability and statistical inference in the seventeenth and eighteenth centuries, respectively.

Philosophically, the profound changes at that time could be summarized as processes of splitting and polarization; of externalization and movement toward the surface, toward the visible; and of an uprooting and decentering. Widespread monetization in the preceding centuries had implied that everything could be measured, that the world could be understood in mathematical terms, and that judgment itself might be replaced by calculation. As a conspicuous example of a self-sustaining, self-correcting system, the market lent plausibility to mechanistic theories of nature. The world and our knowledge of it were becoming depersonalized. A concrete sense of locatedness gave way to a newly abstract sense of space and time, without origin or limit. As resemblance was superseded by order and measurement, and interpretation gave way to representation, everything was assimilated to the concept of sign. In that movement, knowledge became externalized, thenceforth something more like data than understanding. Aristotelian causes could not be found by empirical methods, which could only determine presence or absence, the frequency of co-occurrence of signs; the connections between things were lost; entities became detachable from their attributes. As language broke away from things, and acquired a new arbitrariness, the perspective-boundness of knowledge, and the finitude of perspective, became suddenly apparent. Subject emerged as distinct from object, and, in a serious sense, the species became self-conscious. In perhaps its main beneficent influence, the Reformation championed the epistemological authorization of the individual human knower. At the same time, it was vigorously renewing the moral rejection of the body and hence damning its cognitive apparatus as unclean. The body, with its weight of finitude, had to be transcended to attain the uncorrupted vision of pure spirituality, the infinite remove of the angels.

The period has often been referred to, understandably, as one of parturition for the species. It might, with some justification, alternatively be regarded as the threshold of adolescence, with the attainment of critical self-awareness. The way had been philosophically prepared for passage to cognitive maturity, but it was not to be achieved so easily. The epistemological transformations that would have made it possible also had strong psychological side effects. The loss of connection to the world, a sense of the superficiality and tenuousness of knowledge, the loss of a sense of belonging and of place, the new awareness of finitude and epistemological aloneness—all must have entailed on some level acute anxiety and insecurity and an ensuing frantic search for a solid footing.

In the midst of that trauma, the formation of the modern dualistic concept of probability at once gave expression to the crisis of knowledge—to the externality, the superficiality of knowledge, the supersession of substantial connections by mere frequency counts—and provided the key to its solution, in the mathematical formalization of all empirical knowledge by statistical inference. The giddy prospect of quantifying inference effectively short-circuited the path to intellectual maturity. So powerful was that promise that in the twenty-first century it would still be widely assumed that the problems with statistical inference were of merely a technical nature.

Underlying and reinforcing these developments was the set of defensive reactions engendered by the moral rejection of the body. Since the goal of a ruthless subjugation of the body, a rigid control of the passions and the appetites, was ultimately neither realistic nor appropriate, the result could only be splits, polarizations, and projections on a major scale. Among these were sharp polarizations with respect to age and gender. In the former case, the disowning of childhood dependency, by exaggerating adult needs for control, contributed to the formation and maintenance of the subject-object dichotomy in its various forms and of the modern ideal of objectivity. In the latter case, women were aligned with the scorned and feared body and nature, men with exalted reason and science. That that value-laden polarization is still strong today is apparent in the glorification of “hard” data and rigorous experiments and the contempt for approaches that are “soft,” intuitive, or “touchy-feely.” Statistical inference and quantification more generally are the proudest claim of the social sciences to objectivity and to masculine self-esteem. These identifications should also tell us something about the obstacles to change.

References

- Ableman, P. (1985). *Beyond nakedness*. Los Angeles: Elysium Growth Press.
- Acree, M. (2005). Who's your daddy? Authority, asceticism, and the spread of liberty. *Liberty*, 19(4), 26–32.
- Albury, W. R. (1986). The order of ideas: Condillac's method of analysis as a political instrument in the French Revolution. In J. A. Schuster & R. R. Yeo (Eds.), *The politics and rhetoric of scientific method: Historical studies* (pp. 203–225). Dordrecht, The Netherlands: Reidel.
- Andreski, S. (1982, May). The syphilitic shock. *Encounter*, 57(5), 7–26.

- Andreski, S. (1989). *Syphilis, Puritanism, and witchcraft: Historical explanations in the light of medicine and psychoanalysis, with a forecast about AIDS*. New York, NY: St. Martin's Press.
- Arbuthnott, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27, 186–190.
- Ariès, P. (1960). *L'enfant et la vie familiale sous l'ancien régime* [The child and family life under the ancien régime; translated as *Centuries of childhood*]. Paris, France: Plon.
- Ariès, P. (1989). Introduction. In R. Chartier (Ed.), *A history of private life: Vol. 3. Passions of the Renaissance* (pp. 1–11). Cambridge, MA: Belknap Press.
- Arnauld, A. (1662). *La logique, ou l'art de penser* [Logic, or the art of thinking]. Paris: Jean de Launay.
- Arnauld, A. (1683). *La logique, ou l'art de penser* [Logic, or the art of thinking] (5th ed.). Paris: G. Despres.
- Bacon, F. (1627). *Sylva sylvarum, or a naturall history in ten centuries*. London: William Rawley.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Benjamin, J. (1988). *The bonds of love: Psychoanalysis, feminism, and the problem of domination*. New York, NY: Pantheon.
- Berman, M. (1984). *The reenchantment of the world*. New York, NY: Bantam.
- Berman, M. (1989). *Coming to our senses: Body and spirit in the hidden history of the West*. New York, NY: Simon and Schuster.
- Bernstein, B. (1971). *Class, codes, and control*. London, UK: Routledge, Kegan, Paul.
- Bonner, A. (Ed. and trans.) (1985). *Selected works of Ramon Llull* (Vol. 1). Princeton, NJ: Princeton University Press.
- Bordo, S. R. (1983). The flight to objectivity: Essays on Cartesianism and culture (Doctoral dissertation, State University of New York at Stony Brook, 1982). *Dissertation Abstracts International*, 43, 3621A. (University Microfilms No. 83-07390).
- Bordo, S. R. (1987). *The flight to objectivity: Essays on Cartesianism and culture*. Albany, NY: State University of New York Press.
- Braunstein, P. (1988). Toward intimacy: The fourteenth and fifteenth centuries. In G. Duby (Ed.), *A history of private life: Vol. 2. Revelations of the medieval world* (pp. 535–630). Cambridge, MA: Belknap Press.
- Buroker, J. V. (1993). The Port-Royal semantics of terms. *Synthèse*, 96, 455–475.
- Buss, A. R. (Ed.). (1979). *Psychology in social context*. New York, NY: Irvington.
- Caput, J. P. (1972). *La langue française: Histoire d'une institution* [The French language: History of an institution]. Paris, France: Larousse.
- Cashman, T. M. (1974). Man's place in nature according to Blaise Pascal (Doctoral dissertation, Columbia University, 1974). *Dissertation Abstracts International*, 35, 3805A.
- Chan, W.-T. (Ed.). (1963). *A source book in Chinese philosophy*. Princeton, NJ: Princeton University Press.
- Cohen, H. (1980). *Equal rights for children*. Totowa, NJ: Littlefield, Adams.
- Collingwood, R. G. (1956). *The idea of history*. London, UK: Oxford University Press.
- Condillac (E. B. de) (1822). *La logique, ou les premiers développemens de l'art de penser* [Logic, or the first steps in the art of thinking]. In *Oeuvres complètes de Condillac* [Complete works of Condillac] (Vol. 15, pp. 317–496). Paris, France: Lecoïnte et Durey & Tourneux. (Original work published 1780).
- Contamine, P. (1988). Peasant hearth to papal palace: The fourteenth and fifteenth centuries. In G. Duby (Ed.), *A history of private life: Vol. 2. Revelations of the medieval world* (pp. 425–505). Cambridge, MA: Belknap Press.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684.
- Crosby, A. W. (1997). *The measure of reality: Quantification and Western society, 1250–1600*. Cambridge, UK: Cambridge University Press.

- Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.
- Daston, L. (1992). Objectivity and the escape from perspective. *Social Studies of Science*, 22, 597–618.
- Dawson, C. (1950). *Religion and the rise of Western culture*. New York, NY: Sheed & Ward.
- de la Roncière, C. (1988). Tuscan notables on the eve of the Renaissance. In G. Duby (Ed.), *A history of private life: Vol. 2. Revelations of the medieval world* (pp. 157–309). Cambridge, MA: Belknap Press.
- de Mause, L. (1988). The evolution of childhood. In L. de Mause (Ed.), *The history of childhood: The untold story of child abuse*. New York: Peter Bedrick Books. (Original work published 1974).
- de St. Exupéry, A. (1999). *Le petit prince*. Paris, France: Gallimard. (Original work published 1943).
- Diringer, D. (1968). *The alphabet: A key to the history of man* (3rd ed.). London, UK: Hutchinson.
- Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge, MA: Harvard University Press.
- Dreyfus, H. L. (1979). *What computers can't do* (rev. ed.). New York, NY: Harper Colophon Books.
- Duby, G. (1988a). Private power, public power. In G. Duby (Ed.), *A history of private life: Vol. 2. Revelations of the medieval world* (pp. 3–31). Cambridge, MA: Belknap Press.
- Duby, G. (1988b). Solitude: Eleventh to thirteenth century. In G. Duby (Ed.), *A history of private life: Vol. 2. Revelations of the medieval world* (pp. 509–533). Cambridge, MA: Belknap Press.
- Easlea, B. (1981). *Science and sexual oppression: Patriarchy's confrontation with woman and nature*. London, UK: Weidenfeld and Nicolson.
- Erdmann, J. E. (1893). *A history of philosophy: Vol. I. Ancient and mediæval philosophy* (3rd ed.). London: Swan Sonnenschein.
- Faludi, S. (1991). *Backlash: The undeclared war against American women*. New York: Crown.
- Farson, R. (1974). *Birthrights*. New York, NY: Macmillan.
- Fischer, D. H. (1970). *Historians' fallacies: Toward a logic of historical thought*. New York, NY: Harper Torchbooks.
- Flax, J. (1983). Political philosophy and the patriarchal unconscious: A psychoanalytic perspective on epistemology and metaphysics. In S. Harding & M. B. Hintikka (Eds.), *Discovering reality: Feminist perspectives on epistemology, metaphysics, methodology, and philosophy of science* (pp. 245–281). Dordrecht, The Netherlands: Reidel.
- Flax, J. (1990). *Thinking fragments: Psychoanalysis, feminism, and postmodernism in the contemporary West*. Berkeley, CA: University of California Press.
- Foucault, M. (1973a). *Madness and civilization: A history of insanity in the Age of Reason*. New York, NY: Vintage.
- Foucault, M. (1973b). *The order of things: An archaeology of the human sciences*. New York, NY: Vintage.
- Francastel, P. (1977). *Peinture et société: Naissance et destruction d'un espace plastique de la renaissance au cubisme* [Painting and society: Birth and destruction of a plastic space from the Renaissance to cubism]. Paris, France: Denoël/Gonthier. (Original work published 1951).
- Franklin, J. (2001). *The science of conjecture: Evidence and probability before Pascal*. Baltimore, MD: Johns Hopkins University Press.
- Friedman, D. (1996). *Hidden order: The economics of everyday life*. New York, NY: Harper.
- Friedman, N. (1967). *The social nature of psychological research*. New York, NY: Basic Books.
- Garber, D., & Zabell, S. (1979). On the emergence of probability. *Archive for History of Exact Sciences*, 21, 33–53.
- Gardner, H. (1985). *The mind's new science*. New York, NY: Basic Books.
- Gardner, M. (1958). *Logic machines and diagrams*. New York, NY: McGraw-Hill.
- Gebser, J. (1985). *The ever-present origin*. (Trans. N. Barstad). Athens, Greece: Ohio University Press. (Original work published 1949–1953).
- Glanvill, J. (1885). *Scepsis scientifica: Or, confess ignorance, the way to science; in an essay of the vanity of dogmatizing and confident opinion* (J. Owen, Ed.). London, UK: Kegan, Paul, Trench & Co. (Original work published 1665).

- Graham, L., Lepenies, W., & Weingart, P. (Eds.). (1983). *Functions and uses of disciplinary histories*. Dordrecht, The Netherlands: Reidel.
- Griffiths, W. J., Jr. (1962). Effects of stress on susceptibility of albino rats to sound-induced convulsions. *Psychological Reports*, 11, 663–665.
- Hacking, I. (1975). *The emergence of probability*. Cambridge, UK: Cambridge University Press.
- Hadden, R. W. (1994). *On the shoulders of merchants: Exchange and the mathematical conception of nature in early modern Europe*. Albany, NY: State University of New York Press.
- Havelock, E. A. (1963). *Preface to Plato*. Cambridge, MA: Belknap Press.
- Hunter, M. (1981). *Science and society in Restoration England*. Cambridge, UK: Cambridge University Press.
- Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Boston: Houghton Mifflin.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press. (1st ed., 1939).
- Johnson, M. (1991). Knowing through the body. *Philosophical Psychology*, 4, 3–18.
- Joseph, H. W. B. (1916). *An introduction to logic* (2nd ed.). Oxford, UK: Clarendon Press.
- Kahn, C. H. (1979). *The art and thought of Heraclitus: An edition of the fragments with translation and commentary*. Cambridge, UK: Cambridge University Press.
- Kassler, J. C. (1986). The emergence of probability reconsidered. *Archives Internationales d'Histoire des Sciences*, 36, 17–44.
- Kaye, J. (1998). *Economy and nature in the fourteenth century: Money, market exchange, and the emergence of scientific thought*. Cambridge, UK: Cambridge University Press.
- Kaye, J. B. (1991). Quantification of quality: The impact of money and monetization on the development of scientific thought in the 14th century. *Dissertation Abstracts International*, 52, 2675A. (University Microfilms No. 92-00353).
- Kearns, E. J. (1979). *Ideas in 17th-century France: The most important thinkers and the climate of ideas in which they worked*. Manchester, UK: University of Manchester Press.
- Keller, E. F. (1983). *A feeling for the organism: The life and work of Barbara McClintock*. New York, NY: Holt.
- Keller, E. F. (1985). *Reflections on gender and science*. New Haven, CT: Yale University Press.
- Keller, E. F., & Grontkowski, C. R. (1983). The mind's eye. In S. Harding & M. B. Hintikka (Eds.), *Discovering reality: Feminist perspectives on epistemology, metaphysics, methodology, and philosophy of science* (pp. 207–224). Dordrecht, The Netherlands: Reidel.
- Kendall, M. G. (1960). Studies in the history of probability and statistics. X. Where shall the history of statistics begin? *Biometrika*, 47, 447–449.
- Kendall, M. G. (1963). Studies in the history of probability and statistics. XIII. Isaac Todhunter's *History of the Mathematical Theory of Probability*. *Biometrika*, 50, 204–205.
- Keynes, J. M. (1973). *A treatise on probability*. New York, NY: St. Martins Press. (Original work published 1921).
- Kneale, W. (1949). *Probability and induction*. Oxford, UK: Clarendon Press.
- Knight, C. (1991). *Blood relations: Menstruation and the origins of culture*. New Haven, CT: Yale University Press.
- Koestler, A. (1972). *The roots of coincidence*. New York, NY: Random House.
- Kubovy, M. (1986). *The psychology of perspective and Renaissance art*. Cambridge, UK: Cambridge University Press.
- Liddell, H. G., & Scott, R. (1968). *A Greek-English lexicon*. Oxford, UK: Clarendon Press. (1st ed., 1843).
- Lucas, R. V. (1988). Hic Mulier: The female transvestite in early modern England. *Renaissance and Reformation*, 24, 65–84.
- Maclean, I. (1998). Foucault's Renaissance episteme reassessed: An Aristotelian counterblast. *Journal of the History of Ideas*, 59, 149–166.
- Maclean, I. (2002). *Logic, signs and nature in the Renaissance: The case of learned medicine*. Cambridge, UK: Cambridge University Press.
- Mahalanobis, P. C. (1957). The foundations of statistics. *Sankhya*, 18, 183–194.

- Mailer, N. (1966). *Cannibals and Christians*. New York, NY: Dial Press.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, The Netherlands: Reidel.
- McCall, S. (Ed.). (1967). *Polish logic, 1920–1939*. Oxford, UK: Clarendon Press.
- McDonough, R. (1991). Plato's not to blame for cognitive science. *Ancient Philosophy*, 11, 301–314.
- Merchant, C. (1980). *The death of nature: Women, ecology, and the Scientific Revolution*. San Francisco, CA: Harper and Row.
- Mill, J. S. (1846). *A system of logic, rationicaive and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. New York, NY: Harper & Brothers.
- Mishler, E. G. (1979). Meaning in context: Is there any other kind? *Harvard Educational Review*, 49, 1–19.
- Montagu, B. (Ed.). (1850). *The works of Francis Bacon*. Philadelphia, PA: A. Hart, Late Carey & Hart.
- O'Neill, J. (1991). *Worlds without content: Against formalism*. London, UK: Routledge.
- Onians, R. B. (1951). *The origins of European thought about the body, the mind, the soul, the world, time, and fate*. Cambridge, UK: Cambridge University Press.
- Pearson, K. (1978). *The history of statistics in the 17th and 18th centuries, against the changing background of intellectual, scientific, and religious thought* (E. S. Pearson, Ed.). London, UK: Griffin.
- Pinol-Douriez, M. (1975). *La construction de l'espace* [The construction of space]. Neuchâtel, Switzerland: Delachaux et Niestlé.
- Porter, T. M. (1992a). Objectivity as standardization: The rhetoric of impersonality in measurement, statistics, and cost-benefit analysis. *Annals of Scholarship*, 9, 19–59.
- Porter, T. M. (1992b). Quantification and the accounting ideal in science. *Social Studies of Science*, 22, 633–652.
- Pusey, E. B. (Trans.) (1966). *The confessions of St. Augustine*. London, UK: Dent.
- Pye, M. (2015). *The edge of the world: A cultural history of the North Sea and the transformation of Europe*. New York, NY: Pegasus Books.
- Radding, C. M. (1985). *A world made by men: Cognition and society, 400–1200*. Chapel Hill, NC: University of North Carolina Press.
- Radhakrishnan, S., & Moore, C. A. (Eds.). (1957). *A sourcebook in Indian philosophy*. Princeton, NJ: Princeton University Press.
- Reichenbach, H. (1949). *The theory of probability* (2nd ed.). Berkeley, CA: University of California Press.
- Richards, E., & Schuster, J. (1989). The feminine method as myth and accounting resource: A challenge to gender studies and social studies of science. *Social Studies of Science*, 19, 697–720. (with discussion).
- Rothbard, M. (2011). *Conceived in liberty*. Auburn, AL: Mises Institute. (Original work published 1979).
- Rüstow, A. (1980). *Freedom and domination: A historical critique of civilization*. Princeton, NJ: Princeton University Press.
- Rychlak, J. F. (1991). *Artificial intelligence and human reason: A teleological critique*. New York, NY: Columbia University Press.
- Schachtel, E. G. (1959). *Metamorphosis*. New York, NY: Basic Books.
- Schuster, J. A., & Yeo, R. R. (Eds.). (1986). *The politics and rhetoric of scientific method: Historical studies*. Dordrecht, The Netherlands: Reidel.
- Shapiro, K. J. (1985). *Bodily reflective modes: A phenomenological method for psychology*. Durham, NC: Duke University Press.
- Sheynin, O. B. (1974). On the prehistory of the theory of probability. *Archive for History of Exact Sciences*, 12, 97–141.

- Sheynin, O. B. (1980). On the history of the statistical method in biology. *Archive for History of Exact Sciences*, 22, 323–371.
- Sheynin, O. B. (1982). On the history of medical statistics. *Archive for History of Exact Sciences*, 26, 241–286.
- Solomon, R. C. (1990). *A passion for justice: Emotions and the origins of the social contract*. Reading, MA: Addison-Wesley.
- Sylla, E. D. (2006). *The art of conjecturing*. Baltimore, MD: Johns Hopkins University Press.
- Todhunter, I. (1865). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge, UK: Macmillan.
- Tough, J. K. (1977). *The development of meaning: A study of children's use of language*. New York, NY: Halsted Press.
- Van Creveld, M. (1999). *The rise and decline of the state*. Cambridge, UK: Cambridge University Press.
- Van den Daele, W. (1977). The social construction of science: Institutionalisation and definition of positive science in the latter half of the seventeenth century. In E. Mendelsohn, P. Weingart, & R. Whitley (Eds.), *The social production of scientific knowledge* (pp. 27–54). Dordrecht, The Netherlands: Reidel.
- Van Hée, L. (1926). The *Ch'ou-jen Chuan* of Yüan Yüan. *Isis*, 8, 103–118.
- von Fritz, K. (1974). *Nous, noein* and their derivatives in pre-Socratic philosophy (excluding Anaxagoras). In A. P. D. Mourelatos (Ed.), *The pre-Socratics* (pp. 23–85). Garden City, NY: Anchor Press.
- Werner, H., & Kaplan, B. (1963). *Symbol formation: An organismic-developmental approach to the psychology of language*. New York, NY: Wiley.
- Windelband, W. (1958). *A history of philosophy*. New York, NY: Harper Torchbooks. (Original work published 1901).
- Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex Publishing.

Chapter 3

Origin of the Modern Concept of Probability

The modern dualistic concept of probability was born with the assimilation of the mathematical calculus of gambling to an older, nonmathematical concept of probability that referred to reliability or trustworthiness. The main task of this chapter, after sketching the development of these concepts in their own right, is to understand how they were brought together. The challenge is an interesting one, just because, as poor a fit as it was, we have come to look on the link as natural and obvious.

3.1 Why Gambling per se Didn't Lead to a Theory of Mathematical Probability

Although playing cards were unknown before 1350 (Hald, 1990), gambling itself dates from antiquity (the English word *aleatory* comes from the Latin word for die; every high school student of Latin remembers Caesar's *Alea jacta est*); so it is curious that throughout that time no one ever attempted (or attempted successfully) to improve his or her luck by calculation; and it will be instructive to consider some possible reasons why the calculus of probability should not have been developed sooner.

One hypothesis points to the fact that sets of equiprobable alternatives, such as we have with a fair die or a fair coin, were either not available or were not treated as such. Brakel (1976) notes that translations from Greek and Latin often give *dice* incorrectly in place of *astralagi*. The latter, one of the commonest gaming devices of ancient Egypt, Greece, and Rome, were made from the ankle bones of hooved animals such as sheep or deer. Roughly a parallelepiped, the astralagus could fall on one of four faces, two of them broad and two narrow. (The other two faces were too rounded for the bone to come to rest on them.) In one of the earliest attempts at the mathematics of gambling, around 1560, Cardano enumerated possible outcomes

from throwing one, two, or three dice; but in treating outcomes with astralagi, he took no account of their asymmetry. His oversight seems just slightly less curious when we note that outcomes were not valued in proportion to their rarity (no doubt good for business, Franklin, 2001, observes). The narrow faces were given values of 1 and 6 and the broad faces values of 3 and 4; and moreover:

The most widely used rule attributed the highest value to the throw which showed different faces for each of the four bones. This throw (1, 3, 4, 6) which was called “Aphrodite” or “Venus” was therefore valued higher than all the other, less frequent throws out of the remaining 34 combinations, including the throw 6, 6, 6, 6, which has a much lower probability. (Sambursky, 1956, p. 45)

It is possible that the first cubical dice may have been formed by filing down astralagi (David, 1955), but dice did not in any event immediately supplant astralagi, since they were also used about as far back. Evidence is mixed on whether ancient dice offered equiprobable sets. Hacking (1975) reports that:

The dice in the cabinets of the Cairo Museum of Antiquities, which the guards kindly allowed me to roll for a long afternoon, appear to be exquisitely well balanced. Indeed a couple of rather irregular looking ones were so well balanced as to suggest that they had been filed off at this or that corner just to make them equiprobable. (p. 4)

David (1955), on the other hand, also reports having thrown many ancient dice and having found them nearly all biased, in various ways.

If fair dice, which would yield similar relative frequencies in long series of tosses, were the exception rather than the rule, then we would have less reason to expect a calculus to be developed. A deeper reason, though, has to do with the use of dice and astralagi in divination. For this purpose, they were sometimes favored over other randomizing procedures, like cutting up fowl. As Hacking (1975) remarks:

Chicken guts are hard to read and invite flights of fancy or corruption. We know that the Israelites, ever sceptical of their conniving priests, preferred the lot whose meaning is open for all to read. Lotteries and dice make a good way of consulting gods directly. (p. 3)

But if we suppose the fall of dice to be determined by the gods, we should be little inclined to scrutinize the outcomes to detect empirical laws. Moreover, as David (1955) suggests:

It would seem a reasonable inference . . . that the mystery and awe which the religious ceremony would lend to the casting of lots for purposes of divination would prevent the thinking person from speculating too deeply about it. Any attempt to try to forecast the result of a throw could undoubtedly be interpreted as an attempt to forecast the action of the deity concerned, and such an act of impiety might be expected to bring ill luck in its train. (p. 8)

As Daston (1988) picturesquely puts it, “Recreational or commercial gambling was therefore condemned as the equivalent of a crank telephone call to the deity” (p. 154).

If the original purpose and meaning of dice throwing, in other words, were to read signs from the gods regarding future events, it begins to seem possible that even regular gamblers would not think at all, like their modern counterparts, in

terms of playing the odds to try to maximize long-run average gain; the intention may not have gone beyond the amusement of the stake. Or, perhaps in a manner not so unlike the amateurs who flock to casinos today, they were seeking signs of whether the gods favored *them*. Schneider (1988) contends that what we have thought of since the seventeenth century as games of chance were theretofore treated as games of skill.¹ Cardano regarded “fortune” as a trait like a skill, even though he recognized that, in attributing a spectacular winning streak of his to his good fortune, rather than to the vagaries of chance, he was rendering his own gambling calculations pointless (Daston, 1992). Modern slot machine enthusiasts are aware that there are such things as mathematical probabilities and that *somebody* knows how to calculate them; they are undoubtedly also aware that the odds always favor the house—and hence that over a sufficiently long run, they are guaranteed to be ruined; but those considerations have little if any bearing on their motivation or experience in gambling. The probability calculus has little relevance, in short, to the purpose either of divination or amusement; it becomes relevant only when gambling is seen as a way of making money.²

But it may be that that idea is a comparatively recent development, perhaps brought about itself by the probability calculus. I note that Lorraine Daston (1987) has made a similar observation about the delayed application of mathematical probability to insurance: In the eighteenth century, she argues, there was at best a vague distinction between insurance and gambling, and even long after mortality tables were available, insurers showed little interest in using statistics to maximize their return: It was information on individuals rather than aggregates that was relevant to assessment of risk.

It may be, in fact, that development of the probability calculus awaited the shift from an agricultural to a commercial economy at the end of the Middle Ages. Schneider (1980) observes that from Aristotle through Kepler, the realm of chance, as the radically contingent, was regarded as inaccessible to science. Since mathematics could thus not have been brought to bear on the subject by traditional academics, he suggests, that development had to await commercial interests. Certain developments in contract law, in particular, had an impact on the perception of gambling. When insurance first appeared, for maritime commerce in the

¹The distinction evidently evolved rather slowly, as it does in ontogenesis. It is odd for us today, for example, to see Bernoulli (1713) applying his calculations of expectation indifferently to card games or tennis. The corresponding differentiation between chance and skill in ontogenesis was studied by Marguerite Loosli-Usteri (1931). She showed children how to make inkblots by scattering ink “at random” over a piece of paper and then folding over the paper. Having seen how the result could be interpreted in Rorschach fashion, many children promptly exclaimed, “I think I'll make an elephant!”—then were disappointed and perplexed when their creations bore no resemblance to an elephant.

²Its relevance even there may be limited. The first textbook on calculations in gambling, by Huygens (see below), was hotly denounced by one prominent gambler who had studied it: He could make enormously more money just by cheating (Daston, 1992a).

Mediterranean ports of the fourteenth century (Schneider, 1988),³ it was assimilated, according to Daston (1987), to the model of gambling. (Indeed, Schneider, 1980, conjectures that the Italians might have proceeded to develop the calculus of gambling well before the seventeenth century, had their work not been interrupted by the Counter Reformation.) Neither insurance nor gambling was evidently regarded so much as a way of making money as a simple, if illicit, amusement. Daston notes that, because of its association with gambling, life insurance was illegal in most European countries until the nineteenth century. A growing value on family responsibility by the eighteenth century then permitted insurance to diverge as something honorable. The initial connection may also have been facilitated by changes in canon law to cover clergy who had been lending money with interest. The new law, according to Schneider (1988), had the effect of rationalizing gambling as a commercial enterprise, a transaction of mutual risk. Thus in some sense, capitalism may actually be to blame for the modern science of probability and statistics, if what allowed mathematical analysis to be brought to bear on gambling was its assimilation to the realm of commercial contracts.

The foregoing considerations as to why the probability calculus took so long to develop thus point up interesting similarities and differences between the ancient world and the modern. The difference is, fundamentally, that the mathematical aspect has become so much a part of our concept that we cannot easily imagine probability or gambling without its presence at least implicitly. The similarity is that that aspect appears to have little or no relevance to most people's everyday experience or practice.

3.2 The Concept of Probability before the Seventeenth Century

It would be natural to expect that, without the mathematical aspect, the medieval concept of probability would have referred simply to degrees of belief or of certainty, but that assumption turns out to be just a little too simple. Such a meaning did exist and became increasingly prevalent, but it was derivative. The original meaning can be discerned from the other surviving descendants of the root *probus*, good: words like *probity*, *probation*, *probe*, *proof*, *approbation*, and *approval* (also *reprove* and *reprobate*) (Partridge, 1966). The infinitive *probare* meant to find good or to cause to find good, hence to test; *probabilis*—and *probable*—accordingly meant testworthy or reliable. This meaning disappeared by the nineteenth century; the fact that it now strikes us as strange attests to the shift. Thus when Defoe spoke of a “probable doctor,” he meant, not someone who was probably a doctor, but a doctor who could be relied upon; and when Gibbon described a report as “probable but

³The Hansa, when they founded schools, curiously did not teach mathematics and were slow to adopt both maritime insurance and the Dutch method of double-entry bookkeeping (Pye, 2015).

undoubtedly false,” he was referring to the trustworthiness of the source (Hacking, 1975).⁴

This root meaning spread, even in antiquity, to cover a range of neighboring meanings. Thus one spoke not only of a probable authority, or probable conduct, or the like, but also, at times, of a probable doctrine. The opinions of wise men were probable—worthy of being believed—just because the men themselves were probable, being experienced in such things (Deman, 1933).⁵

It is not clear, however, that, if the calculus of gambling had been discovered earlier, it would have been called a calculus of probability much before the seventeenth century. Indeed, that labeling waited half a century as it was. Hacking (1975) argues, in a controversial thesis, that further epistemological changes were necessary to open up a conceptual space for the emergence of mathematical probability.

One of these, to which the concept of probability itself contributed reciprocally, was the unification of knowledge. Throughout Western history, knowledge has been dichotomized in one way or another; the Aristotelian distinction, relayed by Aquinas, was between necessary and contingent knowledge. Indeed, the designation of “knowledge” was often reserved for the former, demonstrative truth; contingent knowledge was *aestimatio* (if based on sensation) or *opinio* (if based on argument). Paralleling the distinction between necessary and contingent knowledge was that between the high and low sciences of the Renaissance. The former included mathematics, physics, and astronomy, the latter medicine, alchemy, and astrology.⁶

In the Renaissance, probability was predicated of opinion, not knowledge, and was therefore, according to Hacking, a concept of the low sciences; it was *not* a

⁴ According to Franklin (2001), ‘‘The first known occurrence of the word *probable* in English is in Trevisa’s translation of Higden’s *Polychronicon*, of about 1385. It says of snakes in Ireland: ‘It is more probable [Latin *probabilis*] and more skilful that this lond was from the bygynnyng alwey without such wormes’’ (p. 127). The use of *probable* here seems natural to modern ears, but the unnaturalness of *skilful* reminds us of the risks of imputing meanings to ancient writings.’

⁵ It should also be noted that the concept of probability has had a tainted history since ancient times. Augustine, making his own issues clear, as always, feared that young men would use the concept of probability to seduce other men’s wives (Franklin, 2001). Bartolomé de Medina, of the School of Salamanca, was most responsible for bringing the concept into disrepute, with his doctrine of probabilism. Medina argued that although it is better to be a virgin than to marry, and better to be pious than rich, we are not obligated to hold ourselves to such a standard of perfection; and by the same token, it is permissible to accept any probable opinion (i.e., one attested to by some authority), even though its opposite may be more probable (i.e., held by a higher authority). Pascal, a Jansenist, considered with Fermat one of the founders of modern probability theory, never used the word *probability* except in denouncing Jesuits (Franklin, 2001).

⁶ Brown (1987) draws the distinction in terms similar to that between modern academic and clinical psychology: that the low sciences were distinguished simply by an orientation that was both practical and singular. He insists that their *methods* were not different; but, in the case of clinical psychology, that has remained a live issue up to the present, in the controversy over nomothetic versus idiographic methods (e.g., Holt, 1962; Allport, 1962).

mathematical concept.⁷ Hacking argues that the two categories were strictly separate:

The limit of increasing probability of opinion might be certain belief, but it is not knowledge: not because it lacks some missing ingredient, but because in general the objects of opinion are not the kinds of propositions that can be objects of knowledge. (p. 22)

His claim would appear to be too strong, inasmuch as Aquinas, according to Deman (1933), held that probabilities can represent statements on the way to conclusive demonstration. Indeed, Deman goes on to argue, as a well-known genetic epistemologist was thought to, that ontogeny in this respect recapitulates history: “Historically no less than individually, by the probable one accedes to the certain” (p. 288). But Hacking may be right at least in the sense that once a proposition were demonstrated, and absorbed into *scientia*, it would not have been said that it had attained the limit of probability; probability would simply have become irrelevant and inapplicable at that point.

But it was just in that respect that the old conception of knowledge stood to change, with the decline of scholasticism and the press of the Renaissance humanists for the unification of knowledge. The final step in that process was the abandonment, by seventeenth-century science, of Aristotelian causes and the top of the Baconian pyramid. The upper limit in experimental inquiry was then “moral certainty,” rather than proof.⁸ Once the claims of both the high sciences and the low sciences were ordered along the same continuum, the distinction tended to collapse. Thus Shapiro (1983, p. 43) quotes a letter of Huygens referring to “different degrees of probability . . . as 100,000 to 1 as in geometrical demonstration”—though Daston (1988) notes that Locke was still refusing to identify the upper end of the probability continuum with certainty.

Hacking (1975) believes that the Renaissance naturalization of authority helped to make it possible for probability to provide a unifying continuum for knowledge. In the medieval period, probability had designated the trustworthiness of wise men. With the shift from human to natural authority, it came to refer to the reliability of natural signs. The original guarantor of probability had been the wisdom of the ancients; with natural signs, the source of their probability had to be their regularity of co-occurrence with the thing signified. Thus, to take a stock example of the time, pallor was a probable sign of pregnancy, but it was not perfectly reliable inasmuch

⁷Franklin (2001) might disagree, referring wryly to the

False Decretals, an influential mixture of old papal letters, quotations taken out of context, and outright forgeries put together somewhere in Western Europe about 850. The passage itself may be much older.

A bishop should not be condemned except with seventy-two witnesses. . . a cardinal priest should not be condemned except with forty-four witnesses, a cardinal deacon of the city of Rome without thirty-six witnesses, a subdeacon, acolyte, exorcist, lector, or doorkeeper except with seven witnesses.

It is the world’s first quantitative theory of probability. Which shows why being quantitative about probability is not necessarily a good thing. (pp. 13–14)

⁸The term *moral certainty* appears to have been originated by Jean Gerson, chancellor of the University of Paris, around 1400. Gerson used the concept in reassuring a priest who was worried about being unworthy to celebrate mass because of “nocturnal pollution.” Pius XII was still trying to clarify the concept in 1942 (Franklin, 2001).

as pallor also sprang from other causes. And thus it was also, in the naturalization of signs, that probability came to be associated with frequency. The change was not a simple substitution, for probability retained, along with its new meaning, its original reference to human authority, and so a certain duality was built in. We observe a similar distinction yet today in jurisprudence, between testimony and circumstantial evidence, and the probability of testimony was to remain a major application for the new calculus well into the nineteenth century.

The first work to mark the distinction explicitly between the two types of evidence—the testimony of persons and the testimony, as it were, of things—was, according to Hacking (1975), the Port-Royal *Logic*. Arnauld (1662) wrote:

In order to judge of the truth of an event, and to determine for myself whether to believe it or not to believe it, it is not necessary to consider it abstractly and in itself, as we should consider a proposition in geometry; but it is necessary to attend to all the circumstances which accompany it, internal as well as external. I call internal circumstances those which belong to the fact itself, and external those which pertain to the persons by whose testimony we are led to believe it. This being done, if all the circumstances are such that it never or very rarely happens that like circumstances are accompanied by falsehood, our mind is led naturally to believe that it is true, and it is right to do so, especially in the conduct of life, which does not demand greater certainty than this moral certainty, and which must even rest satisfied in many situations with the greatest probability. (p. 363)

Arnauld has already introduced the criterion of probability of natural signs, of internal evidence: We no longer rely on the stature of authorities, but on the *frequency* with which the signs tell us the truth. If all signs are probable, those with the greatest probability, those most to be trusted, are those with the most law-like regularity of occurrence.

Hobbes had made the connection between signs and frequencies even earlier, in his *Human Nature*, published in 1650:

This taking of signs by *experience*, is that wherein men do ordinarily think, the difference stands between man and man in *wisdom*, by which they commonly understand a man's whole ability or *power cognitive*; but this is an *error*; for the signs are but *conjectural*; and according as they have often or seldom failed, so their *assurance* is more or less; but *never full and evident*: for though a man have always seen the day and night to follow one another hitherto, yet can he not hence conclude they shall do so, or that they have done so eternally: *experience concludeth nothing universally*. If the signs hit twenty times for one missing, a man may lay a wager of twenty to one of the event; but may not conclude it for a truth. (cited in Hacking, 1975, p. 48)

Here, as Hacking says, probability has emerged in all but name; and, in the same movement, to quote Foucault (1973), “Hume has become possible” (p. 60).

Garber and Zabell (1979) believe Hacking overstates the radicalness of seventeenth-century developments. The association of probability with frequency at least up to Hobbes or Arnauld was still vague and qualitative, and that kind of association, as Hacking is aware, goes back to Aristotle (*Rhetorica*, 1357a35):

A Probability is a thing that usually happens; not, however, as some definitions would suggest, anything whatever that usually happens, but only if it belong to the class of the “contingent” or “variable.” It bears the same relation to that in respect of which it is probable as the universal bears to the particular. (McKeon, 1941, p. 1332)

A more striking, and now quantitative, example is provided by Garber and Zabell, from Nicole Oresme in the fourteenth century (whose thinking in other respects, as was noted in Chap. 2, was far ahead of his time):

The number of stars is even; the number of stars is odd. One [of these statements] is necessary, the other impossible. However, we have doubts as to which is necessary, so that we say of each that it is possible. . . . The number of stars is a cube. Now, indeed, we say that it is possible but not, however, probable or credible or likely [*non tamen probabile aut opinabile aut verisimile*], since such numbers are much fewer than others. . . . The number of stars is not a cube. We say that it is possible, probable, and likely. (cited in Garber & Zabell, 1979, pp. 46-47)

Moreover, these authors point out, the distinction between internal and external evidence, rather than originating with Arnauld, goes back to Augustine:

Among signs, some are natural and some are conventional. Those are natural which, without any desire or intention of signifying, make us aware of something beyond themselves, like smoke which signifies fire. . . . Conventional signs are those which living creatures show to one another for the purpose of conveying, in so far as they are able, the motion of their spirits or something which they have sensed or understood. (cited in Garber & Zabell, 1979, p. 42)

In fact, the way that Cicero used *probable* sounds very much like the modern evidential concept: If a “wise man” were to set out to sail a distance of four miles, with a good crew and clear weather, Cicero says, “it would appear probable that he would get there safely” (cited in Garber & Zabell, 1979, p. 45). Finally, they argue, “The ancient writers were very much ‘contemporaries’ with respect to Medieval and Renaissance thinkers; they were read, discussed, advocated, and criticized not as archaic texts, but as living documents” (Garber & Zabell, 1979, p. 37).

Arnauld was surely classically trained, so it is indeed curious that he would have supposed himself to be original, for example, in marking a distinction between internal and external evidence. On the other hand, as Karl Pearson (1978) remarks, it was not the policy of French authors of that period to acknowledge any sources for their work. What can safely be said, evidently, is that the seventeenth century saw renewed interest in and commentary on these concepts, as if they were being rediscovered. And that may well be enough to sustain Hacking’s point. Probability as yet lacked a metric, but it did provide a base for ordered comparisons of the reliability of claims on a newly unified continuum of knowledge, and the connection with relative frequencies had all but been made explicitly. The metric that was soon to appear came from an unrelated field, however, and was not to be assimilated to the concept of probability for another 50 years.

3.3 The Calculus of Expectation

Montmort (1713) may have been the first to date the founding of the mathematical calculus of probability to the correspondence between Pascal and Fermat in 1654. There is naturally some arbitrariness in the attribution, but, given the suddenness of

the emergence of the calculus, perhaps a surprisingly small amount; in any case, the disputed aspects of the history are of little consequence here. The work done prior to that of Pascal and Fermat amounted to the enumeration of sets of outcomes in a few special gaming situations. Galileo, for example, had determined that there were 25 ways of throwing 9 with 3 dice and 27 ways of throwing 10. (The date of his piece is unknown; it was not published until 1718, but Galileo died in 1642.)

Pascal was consulted by Antoine Gombauld,⁹ the Chevalier de Méré, on several gambling problems, one of them the question of how many throws of two dice were required for at least an even chance of a double 6. Méré reasoned (as Cardano had) by proportions: In tossing one die, there are 6 outcomes, and 4 tosses are required for at least an even chance of a 6; hence, there being 36 outcomes with two dice, 24 trials should be required for at least an even chance of a double 6. As Weaver (1963) shows, the principle of proportionality to which he appealed works well enough when the probability of both events is small; but Pascal demonstrated, with calculations like Galileo's that the chance of a double 6 in 24 throws was only about 0.491, and 0.505 for 25 throws. There has been considerable discussion whether the difference between 0.491 and 0.505 (or between 25/52 and 27/52 for Galileo's problem) could have been detected by experience.

Pascal also took up the problem of how to divide the stakes between two players, when playing must be stopped at a point where one needs n games to win and the other needs m , and he corresponded with Fermat over the solution to this and similar problems. The "problem of points," or division of stakes, had obvious roots in commercial contracts and was very widely discussed. Sylla (2003) has surely identified a major source of contention in these discussions: Solutions could be based either on the respective numbers of games already won by the two players or on the number each still needed to win.

Historians of probability are accustomed to judging that only the probabilistic calculation that looks to the future is correct, while one that looks to the past is erroneous, but it makes more sense to consider the actual business situation of players of games of luck: looking to the past is reasonable if only the original players are involved. Looking to the future is appropriate if a new player is expected to buy into the game. Considerations of this sort governed the mathematics of business partnerships before they were applied to the "problem of points" in games. (p. 325)

Schneider (1980) makes the provocative suggestion that these puzzles were really a disguised way of solving problems in merchant banking, direct work on such questions being prohibited by the Church doctrine of the sterility of money. In any event, neither Pascal nor Fermat published on the doctrine of chances, but their calculations did stimulate a young Dutchman to write the first book on the subject. In 1655 Christian Huygens came to Paris, at the age of 26, and heard about their work. He never met Pascal and apparently had trouble getting information from anybody; but, undaunted, he went back to Holland, worked out most of the ideas for himself, and

⁹In the letter describing the meeting, Méré did not name his consultant, and his identity has remained a subject of speculation. Hacking (1975) reports that Mesnard, Pascal's biographer, was unable to find a date on which the men could have met.

wrote them up into a short book. He did correspond with Fermat, who contributed some problems as exercises. Van Schooten, Huygens' professor at Leyden, translated the essay from Dutch into Latin, for a wider audience, and had it published in 1657, under the title *Ratiociniis in Aleae Ludo*, at the end of his *Exercitationum Mathematicarum*. This work remained the standard text until the eighteenth century.

We are today so accustomed to thinking of these problems as involving calculations of probability that it seems strange not to find the word *probability* used in any of the discussions of gambling problems by Pascal, Fermat, or Huygens. The first such use has long been credited to the Port-Royal *Logic*:

There are games where ten persons lay down a crown each, there is only one who gains the whole, and all the others lose: thus each risks only a crown, and may gain nine. If we were to consider only the gain and loss in themselves, it might appear that all have the advantage; but we must consider further that if each may gain nine crowns, and risks losing only one, it is also nine times more probable in relation to each that he will lose his crown, and will not gain the nine. Thus each has for himself nine crowns to hope for, one to lose, nine degrees of probability of losing a crown, and only one of gaining the nine, which puts the matter in a perfect equality. (Arnauld, 1662, pp. 384-385)

Shafer (1976b) suggests that we cannot really tell whether to Arnauld's contemporaries his use of *probability* in a mathematical sense sounded familiar or bizarre, but Franklin (2001) derides Arnauld's¹⁰ use of numbers here as "purely decorative" (p. 362) and offers, from his exhaustive scholarship, some evidence in favor of the former. An unpublished manuscript of Gilles Personne de Roberval, for example, written no later than 1647, used the term *vray semblable* in connection with dice:

I say of a proposition that it is believable and likely (*vray semblable*) when, though not infallible, it has more appearances and signs than its contrary. . . . When there are more signs for one thing than for another, one should conclude for the majority of signs if they are equally considerable. One must believe that one thing will happen rather than another when it has more natural possibilities or when the like has happened more often, as in throwing three dice one should believe, and it is likely, that one will get 10 rather than 4, because 10 can be made up in more ways than 4. (quoted in Franklin, 2001, p. 305)

And even earlier, in 1617, Thomas Gataker, in England, had used the word *probable* in connection with lots. Gataker, challenging the Puritan view that lots were determined by God, asked about the outcome of repeated trials:

Were it certain, yea or probable that they should all light upon the same person? Or were it not frivolous, if not impious, therefore to say, that upon every second shaking or drawing God altereth his sentence, and so to accuse him of inconstancy; or that to several Companies he giveth a several sentence, and so to charge him with contradiction and contrariety? (quoted in Franklin, 2001, p. 285)

Most relevantly, Juan Caramuel, who was said by a contemporary to have had "the ability of eight, the eloquence of five, the judgment of two" (Franklin, 2001, p. 89), referred in 1652 to insurance for shipwrecks in grading probability:

¹⁰Arnauld is the presumptive author. Hacking (1975), inspecting the manuscript in the Bibliothèque Nationale, judged the handwriting in the "probability chapters," at the end of the book, to be different from that in the preceding chapters; but there are no other obvious contenders.

For there are men in their ports called Insurers, who go surety for Sea and Wind; if you pay them five in the hundred, they contract to pay back the full amount, if the ship happens to be wrecked. Therefore in the opinion of these men, of the two propositions, ‘This ship will perish’ and ‘It will not perish,’ the first is probable as five (or perhaps as two or three, for these Insurers profit and grow rich) and the second probable as a hundred. (quoted in Franklin, p. 91)

In light of these examples compiled by Franklin, Arnauld’s use of *probability* does not appear exceptional.

The word which was to become standard for many discussions for over a century was supplied by Van Schooten, in translating Huygens’ *kans* (chance) as *expectation*. Where for Cardano and Galileo the operative concept was something like *facility* (“I can as easily make . . .”), Huygens thought, in the problem of points, or of dividing stakes, in terms of the value of a player’s chance, or prospect, of winning. Freudenthal (1980) has provided a literal translation from the Dutch:

I take it as a fundament . . . that in gambling the chance that somebody has toward something is worth as much as that [with] which, having it, he can arrive at the same chance by an equitable game, that is, where no loss is offered to anybody. For instance: if somebody without my knowledge hides 3 shillings in one hand and 7 shillings in the other and lets me choose which of either I want to get, I say this is worth the same to me as if I had 5 shillings for sure. Because, if I have 5 shillings, I can again arrive at having the same chance to get 3 or 7 shillings, and this by an equitable game. (pp. 114-115)

The modern concept of mathematical expectation is dependent on, or derivative from, probability: If a lottery sells 1000 tickets for a prize of \$1000, we compute the expectation, the fair price of a ticket, as $\$1000/1000 = \1 . Early writers, in contrast, took expectation as fundamental, even, sometimes, after probability emerged as a distinct mathematical concept. Bayes (1763), for example, defined probability in terms of expectation: “The *probability* of any event is the ratio between the value at which an expectation depending upon the happening of the event ought to be computed, and the value of the thing expected upon its happening” (p. 370). In other words, if a lottery offers a single prize of \$1000, and the fair price of a ticket is \$1, then the probability of winning is defined as $\$1/\$1000 = 0.001$.

Definitions like Huygens’ sound circular and pointless to us today, but Daston (1983) suggests that taking expectation as the primitive concept made sense at the time because it was a concept already familiar from the field of law. In contracts covering situations of risk—annuities, maritime insurance, prior purchase of the “cast of a net” of a fisherman, and the like—equal expectation was considered to define an equitable contract. There was generally no basis for calculating expectation; it was evidently understood in the same way as any other market price—in fact, Condorcet was to develop this analogy explicitly. But thus it is, perhaps, that Huygens could say that chances were equal because the game was assumed fair.

Further support for the concept of expectation may have come from Pascal’s wager, which Hacking (1975) points to as the first application of the calculus of gambling to decision-making under uncertainty. Pascal argued, as is well known, that either God exists or He does not, and we may decide to act either as though He exists or does not. The former choice means living a life of piety, which will tend

eventually to make one a believer. If God does not exist, there is only a finite difference in expectation between these two choices—the sacrifice of some earthly pleasures; but if He does exist, the difference between damnation and salvation, being infinite, swamps the difference in the other case; so the prudent individual will wager that God exists.

Pascal's model of reasoning found eager acceptance by a generation which was relinquishing the demonstrative knowledge of scholasticism, but, with a few exceptions like Hobbes, lacked the courage or the conviction of full conventionalism. It suggested itself readily as a model for conduct in all aspects of life. The “reasonable man,” steering a course between the extremes of dogmatism and skepticism, was one who acted always to maximize expectation. The idea that “Probability is the very guide of life” may have been an ancient one—Weaver (1963) attributes it to Cicero—and for that purpose, its original meaning is apt enough. John Wilkins, one of the secretaries of the Royal Society, spoke for the seventeenth century in his posthumously published *Of the Principles and Doctrines of Natural Religion*: “In all the ordinary affairs of life men are used to guide their actions by this rule, namely to incline to that which is most probable and likely when they cannot attain any clear unquestionable certainty” (cited in Hacking, 1975, p. 82). Once probability became, with Bernoulli, an explicitly mathematical concept, it carried the powerful suggestion that ethics could be reduced to mathematics. Of such hopes much of the euphoria of the Enlightenment appears to have been made.

In its role as fundamental to probability, the concept of expectation suffered a curious death. It is curious, in the first place, just because the weaknesses of that conceptual arrangement are so obvious to us now. They came to light at the time in the celebrated Petersburg paradox, so called because Daniel Bernoulli proposed a solution in a memoir to the Academy at St. Petersburg. The problem had been proposed by his cousin Nicholas Bernoulli in a letter to Pierre Montmort, published in the second edition of Montmort's *Essay d'Analyse sur les Jeux de Hazard* (1713, p. 402). Curiously, neither Bernoulli nor Montmort seems to have found the problem particularly interesting, evidently just because the solution was straightforward.

In the Petersburg game, *A* tosses a coin until a head turns up. If *A* gets a head on the first toss, *B* pays *A* \$1 (a crown, in the original). If the first head occurs on the *n*th toss, *A* collects \$ 2^{n-1} . How much should *A* pay *B* to play the game? This question of the fair entrance fee is ordinarily answered (then and now) in terms of the mathematical expectation, which is just the sum of the various payoffs weighted by their probability of occurrence. In this case, *A*'s expectation is $(\frac{1}{2})(\$1) + (\frac{1}{2})^2(\$2) + (\frac{1}{2})^3(\$4) + \dots + (\frac{1}{2})^n(\$2^{n-1}) = \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \dots$. With the expectation being infinite, the disturbing result is that in the long run, *A* is ahead to stake any finite sum—and, in fact, any finite entrance fee would be unfair to *B*. In a culture proclaiming probabilistic expectation as the guide of life, the Petersburg problem precipitated a small crisis: The mathematical prescription could not be farther from what anyone would consider prudent or reasonable, mathematicians included.

Daniel Bernoulli's solution was to introduce a distinction between two kinds of expectation, linked to the distinction between the aggregate and the individual.

Mathematical expectation corresponded to the classical definition, taken from legal contexts, where the goal of justice demanded equal treatment for all and disregard of individual characteristics. Moral expectation, on the other hand, was an economic rather than a legal concept. There the goal of fiscal prudence specifically required taking account of individual circumstances. In particular, Bernoulli proposed, the value of a given sum or of a given risk was proportional to how much we already had:

Bernoulli argued that the plight of a poor man holding a lottery ticket with a $\frac{1}{2}$ probability of winning 20,000 ducats was in no way equivalent to that of a rich man in the same position; the poor man would be foolish not to sell his ticket for 9,000 ducats, although his mathematical expectation was 10,000; the rich man would be ill-advised not to buy it for the same amount. (Daston, 1983, pp. 65-66)

He went on to define a logarithmic relation between wealth and increments in utility, which yielded a finite moral expectation more in line with common sense.

The Petersburg problem certainly undermined the viability of expectation as the fundamental concept in probability theory, and it generated a lively debate (see Jorland, 1987, for a summary of two centuries of discussion of the problem, and Dutka, 1988, for a critique of Jorland); but Daston (1983) argues, interestingly, that it was not so much the Petersburg problem that killed expectation as the fundamental concept, but rather the destruction of the concept on which it depended in turn: the reasonable man. The French Revolution had an impact on its time similar to that of Nazi Germany in our own: The simple optimism of the Enlightenment was no longer possible, and with it went not only some of the enthusiasm for starting the world over from scratch, but also the image of the reasonable man conducting his life by calculation.

The concept of mathematical expectation remained, but in its modern, derivative use. When the new mathematical concept of probability appeared, it would prove more serviceable as the fundamental concept. It could be calculated without reference to expectation, first on the basis of gambling models, later on the basis of empirical frequencies. The latter connection, however, between probability and frequency, was formally made, in the seventeenth century, via the concept of expectation, in the first application of the new calculus outside games of chance: the pricing of annuities. This application also marks the connection of probability with statistics, a link that was not to approach its modern naturalness and importance for another two centuries.

3.4 Statistics

“Statistics” as a plural means to us simply numbers, or more particularly, numbers of things, and there is no acceptable synonym. That usage became standard during the 1830s and 1840s. It seems almost impossible now to talk about such numbers and numerical tables published before that time without using this anachronistic term. That all generations previous to the 1820s managed to get by without it reveals dimly how different was the world

they lived in—a world without suicide rates, unemployment figures, and intelligence quotients. (Porter, 1986, p. 11)

In its original use, statistics had nothing to do with numbers. The term is related, obviously, to the word *state*; and histories usually trace statistics to Gottfried Achenwall, who in 1752 used the (German) word *Statistik* to designate a branch of knowledge relating to statecraft, constitutional history, and the like. In 1798, according to K. Pearson (1978), Sir John Sinclair, in *The Statistical Account of Scotland drawn up from the communications of the ministers of the different parishes*, appropriated the term from the Gottingen school of Achenwall, to mean “an inquiry into the state of a country, for the purpose of ascertaining the quantum of happiness of its inhabitants, and the means of its future improvements” (quoted in Pearson, p. 9). A bitter dispute arose over the title of statistician. Ultimately, Sinclair’s usage prevailed, just because Achenwall’s *Statistik* split into political economy and political philosophy, leaving no adherents to carry the banner.

It was hardly accidental that statistics should have arisen as the science of the state; the discipline and the modern state came into being virtually together. A prerequisite for the modern state, James Scott argues brilliantly in *Seeing Like a State* (1998), is *legibility*. Scott started his inquiry in an attempt to understand why the state was so universally hostile to nomadic peoples; he concluded that legibility is a condition of manipulation; the state requires a concentrated, stable population within easy range. Perception and control of populations—for “the classic state functions of taxation, conscription, and prevention of rebellion” (p. 2)—push for standardization and homogenization. Many devices were instituted for that purpose in the early nineteenth century. The imposition of national languages led to shifts in power, as those who didn’t speak the national language were marginalized. Permanent patronymic surnames were another state invention and imposition, which was convenient to one twentieth-century state in identifying religious backgrounds of citizens. Scott quotes Taine on French education in the nineteenth century: “The Minister of Education could pride himself, just by looking at his watch, which page of Virgil all schoolboys of the Empire were annotating at that exact moment” (p. 219). The consequences of the standardization in agriculture have been particularly disastrous. Scott makes the Hayekian point that knowledge of soils, climate, and growing conditions is highly local. Squanto told settlers to plant corn when oak leaves were the size of a squirrel’s ear; the *Farmer’s Almanac* gives a specific date, which doesn’t have the same portability and must be set late. Wherever variability in conditions, especially in areas of marginal productivity, has favored cultivation of mixed crops, replacement of local knowledge with uniform “scientific” prescriptions has usually been counterproductive. Monoculture is more vulnerable to epidemics, requiring heavier use of pesticides and fertilizer. And, although Scott doesn’t pursue them, the parallel consequences in medicine are obvious enough (e.g., the current crisis with respect to overuse of antibiotics).

Well before the nineteenth- and twentieth-century disasters of making populations, crops, and everything else tractable to statistical analysis, however, the first steps in the new discipline of statistics, as with probability, were taken without the

name it now carries. Early in the fourteenth century, Pearson (1978) notices, Giovanni Villani used church records to make population estimates in Florence, but he failed to start a trend. Counts were sometimes kept, even in medieval times, of the number of taxpayers or of the number of citizens able to bear arms, but these compilations were not used for anything beyond themselves. After the severe plague of 1603, the city of London began keeping weekly bills of mortality¹¹—the first such tally, evidently, counting all persons equally—but the purpose was merely to keep track of the plague, particularly to warn the wealthy when they should move to the country.

In 1662, however, the same year as the Port-Royal *Logic*, John Graunt, having been inspired by Bacon, became the first to suggest, in his *Natural and Political Observations mentioned in a following Index, and made upon the Bills of Mortality*, that these weekly bills might be useful for wider purposes. Hacking (1975) sees as epistemologically significant his use of population figures precisely as *data*, to help assess various theories of the plague. The controversy over the miasmic versus the infection theory would not finally be solved for another two centuries, but Graunt believed his data supported the former: Only the weather, he thought, was variable enough to explain the great fluctuations in plague deaths from week to week.

Graunt drew many other inferences from his data, beyond assessing theories of the plague. He noted that although christenings decreased “by the dying and flying of Teeming-Women”¹² (quoted in K. Pearson, 1978, p. 37), losses from the plague were replaced within 2 years by emigrations from the country; and he used the mortality tables to estimate the population of London. That city, he said, growing in population three times as fast as the rest of the country, was becoming too big a head for the body. Some of his calculations were wildly off, but he still had some sophistication about his data. He was aware, for example, that deaths from socially stigmatized diseases were underreported, as they are today. The cause of death was reported by two women in each parish who were known as body-searchers; with respect to syphilis—which the English disowned even in nomination, as the French pox—he found that they would revise their estimates upward “after the mist of a cup of Ale” (quoted in Pearson, 1978, p. 44).¹³

¹¹David (1962, p. 101) reprints the Bill of Mortality for 1665; among the diseases and casualties, it lists the following incidences: Blasted 5; Burnt and Scalded 8; Distracted 5; Executed 21; Frightened 23; Grief 46; Griping of the Guts 1288; Hanged and made away themselves 7; King's Evill 86; Lethargy 14; Murdered and Shot 5; Plague 68,596; Rising of the Lights 397; Stopping of the Stomach 332; and Teeth and Worms 2614.

¹²I am not sure whether “dying and flying” was an epithet of the day for their joining the company of the angels or whether Graunt was referring merely to their flight to the country.

¹³There has been speculation that Graunt’s book was written by William Petty, starting with Petty’s own claim, after Graunt’s death, to have written it. From Pearson (1978), however, one gets the impression of a man of more ambition than intelligence. Petty was ahead of his time in suggesting various offices and functions—a census, a land registry office, and so on—mostly as positions for himself; but he was also the author of a cubic law of proportion according to which, if a horse weighs as much as 1728 mice, it should be as strong as 12 (Pearson, 1978). It is a little hard, as

Graunt was unacquainted with the work of Huygens or Pascal and did not make use of the concept of expectation. That step was very shortly to be taken, however. In 1671 Jan de Witt proposed to the government of Holland the idea of using mortality tables to determine the price of annuities. Annuities had been used at least since ancient Rome as a way for governments to raise money; but, prior to de Witt, it had not even occurred to Newton to make the price of annuities dependent on the age of the annuitant. Rome apparently set its prices so as virtually to assure itself a handsome gain; England, on the other hand, lost enormous sums through the sale of annuities (Hacking, 1975).

Dutch annuities were paid at 6-month intervals. De Witt calculated that chances of death should be distributed equally over the first 100 half years, then incremented at a certain rate in discrete intervals. His correspondent, John Hudde, the mayor of Amsterdam, disagreed, and thought the chances should be assigned uniformly over the whole lifespan. Either system is startling to us today; but given infant and child mortality in the seventeenth century, both were realistic, and Hudde's scheme was actually good enough.

What is most interesting about this first application of the calculus of gambling is that the connection with empirical data was not made in the way we might well have expected. Today, actuarial statistics stand as a kind of paradigm case of probabilities estimated from empirical frequencies. We have already seen, moreover, that a strong, informal link between probability and frequency was coming to be articulated at the time, for example in Hobbes. Yet there was no talk, in Hudde or de Witt, of measuring probabilities by frequencies—nor even any talk of probabilities. Instead, they set about to partition the human lifespan into units demarcating equal chances of death, in analogy to the partition of outcomes having equal chances in games with dice; and, by what one is tempted to call chance, the scheme worked. If the chances of death in the 6-month intervals given naturally by Dutch annuities had been too discrepant, it is not clear whether they would have grasped the idea of explicit measurement of chances by frequencies, or whether the model would have been abandoned as inapplicable.

Probability and frequency were not to be linked explicitly until the end of the century, by James Bernoulli.¹⁴ The concept of probability figured prominently in all the learned discourse of the day, it seems, except the mathematical. Leibniz was the principal exception, after Arnauld. He believed that probability could be made

Pearson suggests, to imagine as the author of Graunt's book a man who would set a horse to pull against 12 mice.

¹⁴The Bernoullis could be regarded as the Bachs of probability theory. James—also known as Jacob and Jacques—his brothers John and Nicholas, and their respective sons Daniel and Nicholas all made contributions to the field. James, remarkably, was self-taught in mathematics, his father having insisted that he study theology instead. We might suppose that the elder Bernoulli would have been satisfied to read that his son had died of “*une fièvre éthique*” (Anonymous, 1706, p. 134), but in fact *éthique* in this context is a variant of *étique*, emaciated, characteristic of *étisie* (phthisis or tuberculosis). It derives from the Greek *ectikos* rather than *ethikos* and is a variant of *hectique*, feverish. The metaphorical meaning of *hectic* as characterizing a feverish frenzy is a twentieth-century development.

numerical, but his interest, deriving from law, was wholly epistemic and his quantification rudimentary. Although his particular approach to the subject was unique, his ideas testify to the growing prevalence of a quantitative conception of probability. Still, Shafer (1978) believes that any direct influence of Leibniz on Bernoulli in this regard must have been scant; Arnauld's (1662) treatment was a more likely inspiration.

Hald (1990) claims that the seventeenth century had a concept of probability pertaining to events, and that Bernoulli's (1713) revolutionary contribution was to apply the concept to propositions. But we have already seen Caramuel speaking, a century and a half earlier, of mathematical probabilities of propositions, based on insurance contracts. I believe that the chronology is more nearly the reverse. The established concept of probability was epistemic, and, though its source was no longer so much human authority, it still applied to arguments or propositions. Arguably there was a *concept* of probability applicable to events, but linguistically it was expressed either through *chance* or through the related concept of odds. On the other hand, if a proposition specified the outcome in a game of chance (e.g., an ace will be drawn), it was not too great a stretch to speak of the probability of the event itself. This usage became common after Bernoulli (1713) and de Moivre (1718).

When calculations were to be done, in any event, the stand-in for the concept of probability was chance, as in “the chances of death” or “the chance of an ace,” or odds, as in Hobbes’ wager. The next three centuries of epistemology might have been radically different if one of these serviceable concepts had remained the operative word. But the ground was too well prepared; the neighboring concept of probability was too close, and too rich in implications, for the connection to go unmade. Bernoulli, it might be said, was simply naming what was already happening.

3.5 The Art of Conjecturing

The title of the *Ars Conjectandi* is modeled on the Latin title of the Port-Royal *Logic*, *Ars Cogitandi*, the art of thinking. The book is divided into four parts. The first is an explication of Huygens’ brief text, with Bernoulli’s own solutions to the problems. Parts II and III extend the theory and pose additional problems. The traditional nature of these early parts is attested by the fact that the word *probabilitas* does not appear. It is Part IV, “On the Use and Application of the Preceding Doctrines to Civil, Moral, and Economic Matters,” that justifies the title of the book, but Bernoulli never got very far with these applications. The book ends with the proof of what he referred to as his “golden theorem,” a generalization of which Poisson (1837) was to christen the law of large numbers. It is virtually for that theorem alone that Bernoulli has been remembered.

Bernoulli worked on the book during the 1680s (Shafer, 1978), but evidently was never satisfied with it. After his death in 1705, his widow and son Nicholas withheld

the manuscript for fear of plagiarism; Nicholas eventually arranged for it to be published by the Thurneysen brothers in 1713.¹⁵

Whereas the first three parts had been concerned with expectations in games of chance, in Part IV the focus shifts to probability and its role as a guide to rational conduct. It was to be a major contribution of Bernoulli's to draw these two lines together, but he first attempted to develop a quantitative conception of probability apart from games of chance. This aspect of his thought was almost wholly neglected until the 1970s, when Glenn Shafer began calling attention to its importance for understanding the development of the concept of probability.

3.5.1 *The Law of Large Numbers*

For a series of independent trials of a given kind, where an event has a fixed probability p of occurrence, the binomial formula gives the probability of a certain number, r , of occurrences out of n trials. In the first part of his theorem, Bernoulli showed simply that the binomial probability is greatest where $r = np$, to the nearest integer, hence that the most probable proportion of occurrences is p .¹⁶ In the second part of the theorem, he demonstrated algebraically that, for fixed p and increasing n , the middle terms of the binomial expansion account for an increasing proportion of the total probability, and hence, roughly, that the proportion r/n of occurrences, or “successes,” approaches, in probability, the probability p of a success on a single trial. The term “law of large numbers” was technically applied by Poisson (1837) to a more general form of the theorem which he himself proved, and the probabilistic limits which Bernoulli established for the quantity $r - np$ were also narrowed by later mathematicians, but these details do not matter here.

Bernoulli's Theorem, the “weak law of large numbers,” is, as a mathematical proof, unassailable as far as it goes. Although Bernoulli himself appears to have been clear on what he did and did not accomplish with it, the theorem has been the object of much confusion, in more popular presentations, and of controversial interpretations, in the more technical literature. The theorem assumes p , the probability of success on an individual trial, to be known, exactly, in advance, and then allows probability statements to be made about various outcomes in a sequence of trials. It does not provide any basis for an “inverse” inference from an observed proportion

¹⁵ Until very recently, it was said to be Bernoulli's nephew Nicholas—the same age as his son Nicholas—who had, along with Bernoulli's widow, prevented publication for 8 years. In fact, it was the son who withheld the manuscript from the nephew (Sylla, 2006). Yushkevich (1987) traces the confusion all the way back to Christian Wolff in 1716. The son's concerns about plagiarism were not without ground; Bernoulli's brother Johann sold much of Bernoulli's work to the Marquis de l'Hôpital in Paris, who published it as his own (Sylla, 1998).

¹⁶ Bernoulli's use of fractional notation for probabilities— r/t for p , or expressions like $(4/15)c$ —in place of a single symbol can be taken as evidence that probability was still in process of emerging as a concept in its own right.

of successes in a sequence of trials to an unknown p for an individual trial. The reason no such statement could be made is that p itself, so long as it is regarded as a constant, does not have a probability distribution. Theoretically a probability may take a value between 0 and 1; but Bernoulli's Theorem does not say anything further about possible values for p , and in particular cannot set probabilistic limits on such a value.

It is clear that Bernoulli hoped, through the law of large numbers, to ground the use of relative frequencies for the estimation of probabilities. Shafer (1978) notes, in fact, an apparent reference to Graunt's work on mortality statistics, as an obvious possible inspiration for the law of large numbers. In games of chance, probabilities can be determined a priori from considerations of symmetry; Bernoulli was in fact the first to articulate what Venn (1888) was later to christen the principle of indifference¹⁷.

The numbers of outcomes with dice are known; indeed there are obviously as many as there are sides, of which all are equally likely [*proclives*]; since on account of the similarity of the faces and accordingly the weight of the die there should be no reason, why one of the faces should be more inclined [*pronior*] to fall than another, as would happen, if the faces bore different shapes, or the die should be composed of heavier material in one part than in another. (Bernoulli, 1713, p. 224)

But he went on to emphasize how exceptional such cases were:

Who, pray, among mortals will ever determine, for example, the number of diseases, as of so many alternatives [*casuum*], which can invade the innumerable parts of the human body for however long, and bring death to us; & how much more easily this one than that, plague than dropsy, dropsy than fever, should kill a man, so that from these a guess could be made for the future about the circumstance of life and death? (p. 224)

“But as a matter of fact,” he went on to say:

another way is available to us here, by which we might obtain what we seek; & what it is not permitted to ascertain a priori, [is permitted] in any case *a posteriori*, that is, it can be extracted from an event observed in many similar instances; since it ought to be presumed, from however many cases heretofore have each been able to happen & not happen, how

¹⁷In translating from seventeenth- and eighteenth-century Latin (and French) sources, I have elected to produce what is from the standpoint of modern idiomatic English an overly literal rendition, with the result that a degree of intelligibility is sacrificed for some of the flavor of the discourse of the day. There are stylistic differences, to be sure, between seventeenth-century English, Latin, and French; but they are less than those between seventeenth- and twentieth-century exemplars within languages, particularly in English. The trend over the last three centuries, especially pronounced in American English, has been toward greater informality and simplicity. Sentences have become shorter and simpler in their structure; embellishments such as capitals, italics, and hyphens have been dropped—the last change visible within just the last two generations; and the fact that in English the subjunctive is often marked only in the third-person singular has put it well on the way to extinction. Some of these are the result of a natural press for efficiency; the egalitarian demands of compulsory government education, together with the ascendance of electronic over print media, have surely also exerted a certain pull toward simplicity.

many henceforth will be discerned to have happened & not to have happened in a similar state of affairs. (p. 224)

If Titius is one of 300 men of similar age and constitution, he said, and 200 of them died within 10 years, “You will be able to deduce safely enough, that there are two times more cases, for which & for Titius this debt of nature should be acquitted within the next ten years, than for which it should be possible to exceed this limit” (p. 225).

For the purpose, however, of estimating probabilities *a posteriori* from relative frequencies, an inverse form of the theorem, permitting inference from observation to theory, from relative frequency to p , would be more useful than the direct form Bernoulli proved. Hacking (1975) strongly suggests that it was Bernoulli’s realization that he had not succeeded in what he hoped to prove—an inverted form of the law of large numbers—that kept him from publishing. Bernoulli had discussed an inverse form of the theorem in a letter to Leibniz near the end of his life, in 1703; Leibniz was skeptical (Keynes, 1921/ 1973). Stigler (1986), on the other hand, sees no dissatisfaction in Bernoulli with the philosophical aspects of his work; he is impressed with the mathematical disappointment Bernoulli faced. Bernoulli worked out a numerical example of his theorem, to find how many trials were required to be morally certain (i.e., to achieve a probability of 0.999) that the obtained relative frequency of the event was within 1/50 of its probability on an individual trial. The depressing result was 25,550, and Stigler notes that this is literally where Bernoulli laid down his pen: There were no catalogs of anything that were that large. I find Sylla’s (2006) explanation more plausible than these: that Bernoulli was waiting for Leibniz to send him a copy of de Witt’s tract. Leibniz had acknowledged having a copy of the scarce pamphlet in a letter to Bernoulli, and Bernoulli had repeatedly written to request a copy, which he hoped would provide him with a useful example of estimating probabilities from frequencies.

In any event, Bernoulli’s result was disappointing for the purpose of empirical inquiry: either because the number of observations required was hopelessly large, or because the more common situation, as in Bernoulli’s example of diseases, is to want to estimate probabilities from observed relative frequencies rather than the other way around. If the link between probability and frequency was a unidirectional link, however, and a somewhat hypothetical one, it was still a link; and the idea of relative frequency as a metric for probability powerfully shaped the future of that concept.

Yet the assimilation of probability to games of chance, however natural and inevitable it is for us today, posed several problems. The most important of these were the scale of measurement and rules for the combination of evidence. We shall have a close look at these after briefly considering ambiguities in the metaphysical status of probability.

3.5.2 *The Metaphysical Status of Probability*

Reference to knowledge unified the concept of probability. Bernoulli's definition, while adverting to a ratio, sounds psychological or subjective: "Probability is in fact a degree of certainty [*certitudo*], and differs from it as the part from the whole" (1713, p. 211). Bernoulli's assimilation of chances to the concept of probability was facilitated, Shafer (1978) suggests, by his religious determinism. The source of uncertainty, for Bernoulli and for most of his contemporaries, was the incompleteness of our knowledge. In the real world, there is no objective contingency, because everything is controlled by God. "The locus of conjecture cannot be in things, in which certainty in all respects can be achieved" (Bernoulli, 1713, p. 214). Bernoulli recognized explicitly that contingency, like knowledge itself, is relative to the knower. But the consequence is then that there is no fundamental distinction between propositions relating to outcomes in games of chance and statements of uncertainty in science, law, or everyday life. In the terminology of the following centuries, all probabilities were subjective—albeit in the sense of a *rational* subject.

Given the indirect or metacommunicative character of probabilistic discourse, an objective view became possible at the same time as Bernoulli's "subjectivism." The ambiguity is indeed not unique to probability, but characterizes any domain of discourse about discourse: A major debate of the past century has been whether philosophy itself is about reality or merely about the way people use words. Thus if probabilities say something about propositions and propositions say something about reality, then, if I say the probability of a head with this coin is $\frac{1}{2}$, I may well intend to be saying something about the coin and not (just) about my beliefs. "The best example I can recall of the distinction between judging from the subjective and the objective side, in such cases as these," Venn (1888) tells us:

occurred once in a railway train. I met a timid old lady who was in much fear of accidents. I endeavoured to soothe her on the usual statistical ground of the extreme rarity of such events. She listened patiently, and then replied, "Yes, Sir, that is all very well; but I don't see how the real danger will be a bit the less because I don't believe in it." (p. 157n)¹⁸

In the calculation of probabilities, it makes a difference, in particular, whether we regard events as independent in the objective sense that one event has no causal influence on another or in the subjective sense that one event does not affect our belief about the other. D'Alembert (1767) pressed this point particularly hard. In a series of coin tosses:

Either one must take into account the previous tosses to guess the next toss, or one must consider the next toss independently of the preceding; *these two opinions have their partisans*. In the first case, analysis of the chances leads me to believe that if the preceding tosses have been favorable to me, the next toss will be unfavorable; that, in proportion to how

¹⁸ Compare the story told of Niels Bohr that when asked by a journalist about a horseshoe (purported to bring good luck) hanging over his door, he explained that he of course does not believe in such nonsense but heard that it helped even if one did not believe (Shafir & Tversky, 1992, p. 464).

many tosses I have won, I can bet that I will lose the next, and vice versa. I could thus never say: I am having a streak of bad luck, and I shall not risk the next throw, because I could not say that on account of the previous which went against me; these past throws will on the contrary lead me to expect that the following toss will be favorable to me. In the second case, i.e., if one regards the next throw as entirely isolated from the preceding, one has no reason to guess that the next throw will be favorable rather than unfavorable, or unfavorable rather than favorable; thus one could not direct one's behavior according to the details of fate, of good luck, or of bad luck. (pp. 302-303)

In modern parlance, d'Alembert's first case, the hypothesis of serial dependence, is the gambler's fallacy, and modern mathematicians are somewhat less evenhanded in their treatment of it than was d'Alembert.

If we hold, with Bernoulli, to the view of probability as relative to knowledge, then, as Keynes (1921/ [1973](#)) points out, the conditions for his own theorem will virtually never occur in practice. The law of large numbers assumes a very long sequence of trials with constant probability; yet, as d'Alembert was also fond of arguing, there are few, if any, circumstances where we would not sooner or later revise our probabilities in light of extraordinary outcomes:

Suppose Peter and Paul are playing heads or tails, that it is Peter who is tossing, and that heads occurs ten times in a row (which would already be a lot); Paul will inevitably cry out on the tenth toss that this event is not natural, and that surely the coin had been prepared in such a way as always to turn up heads. Paul thus supposes that it is not in nature that an ordinary coin, made and tossed without trickery, should fall ten times in a row on the same side. If you don't find ten times enough, make it twenty; it will always happen that there is no gambler who does not tacitly know this supposition, that the same result can only happen a certain number of times. (d'Alembert, [1767](#), pp. 284-285)

It would theoretically be possible to revise the standard calculus of probability to allow for constantly changing probabilities in the light of experience. Keynes (1921/ [1973](#)), who was the first modern thinker to return to the original meaning of probability as relative to knowledge, made some efforts in this direction; and we shall find later so-called Bayesians taking a somewhat equivocal position on the issue (Chap. 8).

The modern concept of probability, in sum, joined subjective and objective concepts, and the resulting ambiguity persists to this day. The original, epistemic concept of probability was subjective in the sense of being relative to knowledge. Chances in the calculus of gambling were objective in the sense that they were taken to be statements about dice and not (just) about our beliefs about dice. Bernoulli—characteristically, we might say—vaguely straddled the issue. He recognized that probability varied with the knower; the Deity had no uncertainty at all. But his definition of probability as a proportion of certainty kept it tied to the additive calculus of chances. The ambiguity in the metaphysical status of probability, as subjective or objective, has implications for the arithmetic, which concerned some early thinkers somewhat more than their successors.

3.5.3 One Word, Two Scales

The Scale of Measurement Whether we base probabilities a posteriori on relative frequencies in a series of events or a priori on a model of symmetry, probabilities behave like proportions: Over all possible outcomes in a set, they sum to 1. In either interpretation, a 0 probability of X , as a proportion of cases, must represent its impossibility, or the certainty that not- X . A scale that is anchored in certainty at both ends, however, is totally unsuited to the representation of ignorance and therefore to the operations of knowledge and inference. A concept of probability that is going to be useful in inference, if such a thing is possible—one that is going to represent evidence, belief, knowledge, or opinion—must be able to represent ignorance, the starting point of knowledge, an absence of evidence bearing on a question. (The representation of ignorance, it turns out, has been a major issue in modern Bayesian theory, as we shall see in Chap. 8.) We need a scale where the lower end, $p(X) = 0$, represents an absence of evidence in favor of X rather than its impossibility. On such a scale, in situations where there is little evidence bearing on a question one way or another—a situation hardly unknown either in science or in everyday life—the probabilities of a proposition and its contrary (or the conceivable alternatives) may sum to much less than 1.

Given the “transcendental hold” (Blackburn, 1980) the traditional calculus has on us, it is difficult to conceive of any such alternative, nonadditive scale. The field of legal reasoning makes clear, however, the unsuitability of the traditional scale of proportions for Bernoulli’s “civil, moral, and economic matters.” Our tradition of justice requires, for example, starting from a presumption of innocence on the part of the accused; yet the usual scale offers us no means of representing such a state. Laplace (1816) glibly assumed a probability of $\frac{1}{2}$ for the guilt of the accused, but making guilt an even bet is far from a presumption of innocence. Yet a probability of 0 on the traditional scale would mean the impossibility of guilt, which will not do, either.¹⁹ Or consider two witnesses, judged equal in reliability, who disagree on the guilt of the accused. In the absence of other evidence, the traditional calculus would evidently have to assign a probability of $\frac{1}{2}$ to the hypothesis of guilt—regardless of whether both witnesses were regarded as highly reliable or so unreliable as to make their testimony worthless. A scale representing the amount of evidence in support of a proposition would assign probabilities at or near 0 in the latter case both for the hypothesis of guilt and of innocence. The sum of these two probabilities would approach 1 only as the witnesses approached perfect reliability. Note that this “nonadditive” scale takes account of the “weight” of evidence, which is ignored by the traditional calculus.²⁰

¹⁹ Sylla (1990, p. 40) quotes Voltaire: “As there are half-proofs, that is to say half-truths, it is clear there there [are] half-innocent and half-guilty persons. So we start by giving them a half-death, after which we go to lunch.”

²⁰ I leave the concept of weight here at an intuitive level. When it is formally defined, in a later chapter discussing Shafer’s (1976a) theory of belief functions, it turns out that the additive probabilities of the calculus of chances represent infinite contradictory weights of evidence.

Bernoulli, to his credit, recognized, at least vaguely, that the usual calculus of chances could not be applied when frequency data or symmetry models were unavailable and that another metric was required. He attempted to handle the problem with a distinction between what he called *pure* and *mixed* arguments.²¹

Thus I call *pure*, those which prove the thing in certain cases, in such a way that they prove nothing positive in the others: thus *Mixed*, those which prove the thing in some cases, in such a way as to prove in others the contrary of the thing. Here is an example: If in a crowd of rioters someone were to be stabbed with a sword, and it were established by the testimony of trustworthy men observing at a distance that it were a man in a black cloak who committed the deed, and Gracchus be discovered along with three others to be wearing a tunic of that color, this tunic will be some argument for the murder being committed by Gracchus, but it is mixed, since it proves in one case his guilt, in three cases his innocence, according of course to whether the murder were to have been committed by him or by one of the remaining three; and indeed it could not have been committed by one of these, without Gracchus by the same token being considered innocent. But if however in a subsequent hearing Gracchus were to have paled, this pallor of the face is a pure argument: it clearly proves Gracchus' guilt if it stem from a guilty conscience, but does not, on the other hand, prove his innocence if it arise from something else; it can indeed happen that Gracchus should grow pale of other causes, and yet he be the perpetrator of the murder. (p. 218)

The concept of pure arguments raises directly, however, the question of how they are to be measured. In many of the matters of everyday life to which Bernoulli wished to apply the theory of probability, we have neither relative frequencies nor equiprobable partitions. In such situations, it was not clear how numbers might be assigned—nor is it yet. Bernoulli evidently placed his hopes in the compilation of statistics; he wrote as though compiling weather records, for example, would make possible probabilistic predictions. But in fact the probabilities in weather forecasting even today are in most cases merely subjective summaries of evidence; the standard percentages are merely an 11-point rating scale, and it is no discredit to Bernoulli that he didn't foresee that fantastic invention 250 years later.²²

Bernoulli did not attempt to quantify the probability accorded Gracchus' guilt by his pallor. With his reference to other causes of pallor, and to the other cases proving nothing positive, he appears to be thinking that those various causes of pallor might be enumerated to determine the probability of Gracchus' guilt. But, as Shackle

²¹ Weisberg (2014), in his splendidly titled *Willful Ignorance: The Mismeasure of Uncertainty*, appears to be getting at the same distinction with his distinction between *ambiguity* and *doubt*. He is one of the few authors to discuss this aspect of Bernoulli's work; but he never really develops the concepts systematically or usefully. He also discusses Frank Knight's concepts of risk vs. uncertainty, although Knight didn't follow the distinction to its logical conclusion, either. I owe my knowledge of Knight to the Swedish economist Johan Grip (2009).

²² Probabilistic weather predictions, celebrated by Murphy (1991) as "scientifically sound" (p. 302) and "a precise and unambiguous description of the uncertainty in forecasts" (p. 305), go back only to 1965 in the United States (Monahan & Steadman, 1996). Forecasts prior to that time were categorical ("rain," "no rain"), sometimes hedged ("likely rain"). Despite serious efforts with state-of-the-art equipment, computerized forecasting has had a difficult time catching up to human performance; in 1950 the ENIAC took 24 hours to make a 24-hour prediction (Shuman, 1989).

(1961) points out, the probability of various applicants being hired, for example, depends on their qualifications and not just their number. Similarly, the probability accorded Gracchus' guilt by his pallor cannot depend simply on the number of such possible causes, which would have to be large. Suppose, for simplicity, that there were only two other possible causes of Gracchus' pallor: Either he started internal bleeding or he suddenly realized he had forgotten to take the roast out of the oven before heading to court. But even these two are not disjoint. And, if they were, that would not establish a probability of 1/3 for Gracchus' guilt from the pure argument.

Bernoulli's reference to "cases" on both sides of a pure argument has led some readers to interpret these arguments as mixed. Thus Sylla (1990), for example, attempts to quantify the argument about Gracchus' pallor by stipulating that, when people pale under questioning, there is a 60% chance that they are guilty. But she thereby appears to be counting cases—and the argument, so construed, appears to be mixed rather than pure: We are identifying a set of people who pale under questioning, and partitioning it into 60% who are guilty—and 40% who are not. If there are cases to be enumerated on both sides, then we have a mixed argument. Hence it is not clear that, when there are no cases to count on either side, the assessment of probability can be anything other than a vague, informal appraisal of evidence.

Pure arguments will generally comprise the testimony of persons or, as it were, of things, like Gracchus' pallor; and it is obviously these arguments which will predominate outside the restricted context of games of chance and their analogs. Despite their importance, Bernoulli did not clearly recognize that the concept of a pure argument implicitly redefines the scale of probability. On his own definition of probability as a degree of certainty, the lower end of the scale would logically be maximum uncertainty, or an absence of evidence, rather than certainty of the contrary. Near the beginning of Part IV, however, Bernoulli himself, either forgetting about the pure arguments he was about to define or else simply oblivious to the distinction between implied scales, identified a probability 0 with impossibility: "*Possible* is, what has even a very small part of certainty: *Impossible*, what [has] a null or infinitely small [part]" (p. 211). If he had noticed more clearly the difference in scales implied by the two kinds of arguments, he might have hesitated to label them both *probability*.

The Combination of Evidence If we have two distinct scales—one a scale of proportions, whose probabilities over a set of alternatives necessarily sum to 1, the other a scale of evidence, where probabilities over alternatives may sum to less than 1 if there is little evidence one way or the other—we should expect different rules to apply to the combination of probabilities. We can glimpse the need for different rules by considering the ordinary multiplicative law. In the familiar aleatory calculus, probabilities of independent propositions are multiplied to yield the probability of their all being true. In law, and in everyday reasoning, a case is regarded as stronger the more independent elements there are in its favor. The rule for civil cases is that plaintiffs are required to demonstrate a preponderance of evidence—thus presumably a balance of probability—on their side. But Jonathan Cohen (1977) points out that the product rule for conjunction of propositions would make even

moderately complex cases all but impossible to win. If three separate elements in the argument, or three independent witnesses, were adduced with probabilities, respectively, of 0.8, 0.8, and 0.75, the court would regard the case as rather well established. Indeed, the probability of the conjunction should not be less than that of the weakest component. Yet on a traditional calculation the case would be lost, since $0.8 \times 0.8 \times 0.75 < 0.5$. Bernoulli's formula for combining pure arguments amounts to calculating the probability that at least one of the component arguments holds: $1 - (1-0.8)(1-0.8)(1-0.75) = 0.99$. It is questionable whether the probability of the conjunction is best expressed as the probability that at least one of its component propositions is true, but at least Bernoulli's value of 0.99 in this case accords better with common sense than the conventional multiplicative value of 0.48.

The trickier case is combining pure and mixed arguments, and Bernoulli is on still weaker ground here. If we suppose, with Sylla (1990), that, when people pale under questioning, there is a 60% chance that they are guilty, Bernoulli would argue that, of the 40% of cases on which the pure evidence was silent, we should count the $\frac{1}{4}$ supported by the mixed evidence, and that this product should be added to the 60% supported by the pure evidence. Thus the probability of Gracchus' guilt is $0.6 + 0.4(0.25) = 0.7$ —with the consequence that the total probability over the four suspects is $0.7 + 0.75 = 1.45$. Bernoulli did not make this calculation, but he did explicitly allow for probabilities to sum to greater than 1, despite the fact that a probability greater than 1 cannot be reconciled with his definition of probability as a proportion of certainty:

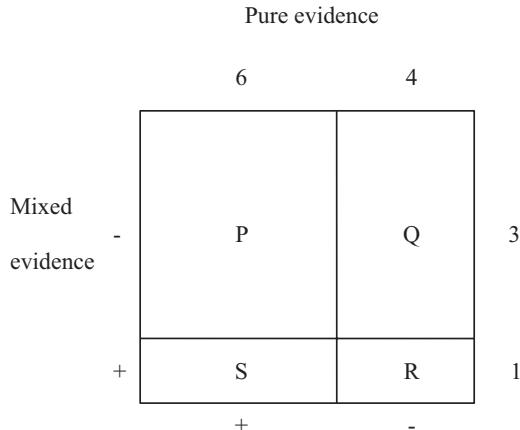
It should be noted that, if arguments, which are to be brought forth on both sides are sufficiently strong, it may happen, that the absolute probability on either side should significantly [*notabiliter*] exceed half of certainty, that is that both of the contraries be rendered probable, though relatively speaking one be less probable than the other; thus it can happen that one have $2/3$ of certainty, while its contrary possesses $\frac{3}{4}$; in that way each of the contraries will be probable, & yet the first less probable than its contrary, and that in the ratio $2/3$ to $3/4$, or 8 to 9. (1713, p. 221)²³

Bernoulli's solution was criticized by Johann Lambert (1764/2002) on the ground that if the number of suspects was very large, so that the mixed evidence for Gracchus' guilt was very small, the pure evidence of pallor would still tend to convict him. We may illustrate Bernoulli's own reasoning with the aid of a graphical

²³ Shafer (1978) commends Bernoulli for not “renormalizing” these probabilities to 8/17 and 9/17. Suppose, he says:

that one pure argument proves 1/10 of the certainty of a thing and another pure argument proves 1/100 of the certainty of the opposite. Then it would seem that the two together, like each singly, fail to prove very much of anything; but a method that renormalizes to force additivity would have it that they together prove 10/11 of the certainty of the thing. (p. 336)

Fig. 3.1 Combining pure and mixed arguments



representation (Fig. 3.1) used by Shafer (1976a). Lambert argued that the cases where the mixed and pure evidence conflict—namely, the rectangle labeled P —should be eliminated from the calculation. Thus the probability of Gracchus' guilt would not be

$$(P + R + S) / (P + Q + R + S) = (18 + 4 + 6) / (18 + 12 + 4 + 6) = 28 / 40 = .7, \quad (3.1)$$

as Bernoulli had calculated, but.

$$(R + S) / (Q + R + S) = (4 + 6) / (12 + 4 + 6) = 10 / 22 \approx .45. \quad (3.2)$$

Lambert generalized Bernoulli's results so that his formulas for pure and mixed arguments drop out as special cases. In the modernized notation of Shafer (1978), for two arguments with probabilities p_1 and p_2 , and contrary probabilities q_1 and q_2 , the probability of the joint assertion is

$$\frac{p_1 + p_2 - p_1 p_2 - p_1 q_2 - p_2 q_1}{1 - p_1 q_2 - p_2 q_1}.$$

The term $p_1 p_2$ corrects the double counting of the overlap in summing p_1 and p_2 , and $p_1 q_2$ and $p_2 q_1$ are the cases of contradictory evidence. If either of the arguments is pure, then the corresponding $q = 0$. For Bernoulli's example of a pure argument having a probability of $2/3$ and a contrary pure argument a probability of $3/4$, Lambert's formula gives a combined probability of $1/3$ and a probability of the contrary of

$$\frac{q_1 + q_2 - q_1 q_2 - p_1 q_2 - p_2 q_1}{1 - p_1 q_2 - p_2 q_1} = \frac{1}{2}$$

Probabilities for Lambert can thus sum to less than 1 but cannot exceed it—consistent with Bernoulli's definition of probability as a proportion of certainty.

Hacking (1974) compared the approaches of Bernoulli and Lambert with reference to an example which usefully illustrates the ways items of evidence can interact:

On a Monday early this April I was on the overnight train from Washington D. C. to Chicago. I had left sun, azaleas and cherry blossom [sic] and was firmly entrenched in thoughts of early summer. In the parlor car I found a scrap of paper saying, "Blizzard due in Chicago tomorrow." The dateline bore the legend, "Monday . . ." but the actual date was torn off. There were two possibilities. Either this was the morning paper, or it was of earlier date. In the latter case, the scrap tells me nothing whatever about tomorrow's weather. It is quite neutral. But in the former case, I may well trust such a categorical forecast of sudden cold. If I were certain the scrap of paper were up-to-date, I could be virtually certain of a cold arrival in Chicago.

The parlor car is kept tidy and my fellow passengers do not seem the sort of people who litter with old scraps of paper. The scrap is clean and fresh. I assess the probability, that the scrap is from today's paper, as 90%. Hence I become 90% certain of a cold arrival. Note that I am not thereby 10% certain of a temperate arrival because the evidence stated so far is simply neutral about the 10% of remaining cases. (p. 114)

Just as Gracchus' pallor was irrelevant if it sprang from his having forgotten his roast, the evidence of the newspaper is irrelevant to tomorrow's forecast if the paper is a week or more out of date. But Hacking also has mixed evidence: From his knowledge of US weather, he assumes the probability of severe cold in early April in Chicago as 20% and thus the probability of temperate weather at 80%. In this case, Bernoulli's formula would give a probability of snow on arrival in Chicago of $0.9 + 0.20(0.10) = 0.92$; Lambert's would give $0.20/(0.20 + 0.80(0.10)) \approx 0.71$. Now suppose, Hacking says, that he is traveling in August, so that the probability of cold weather in Chicago is almost 0. "Then if the scrap of newspaper predicts cold, it *must* be out-of-date. Thus the prediction given in the newspaper may itself be evidence against the newspaper being timely. The paper can testify against itself" (p. 120). The pure argument vanishes. On the other hand, if Hacking were traveling in an area where he were ignorant of weather patterns, then the freshness of the newspaper scrap might well outweigh his vague judgments about the weather. Hacking concluded that Bernoulli's formula makes sense when (but only when) we are certain of our assessments of evidence. But these are just the situations where the evidence cannot conflict and where Bernoulli and Lambert would thus agree. In cases of what has come to be called nonmonotonic reasoning—if learning one piece of evidence prompted us to revise our other probability estimate—then Lambert's formula is more reasonable.

What ultimately matters, however, is not the differences between Bernoulli and Lambert, but the insights—and the mistakes—that they shared. They were both right in recognizing, if only imperfectly, a distinction between two kinds of evidence. Only in situations where the totality of evidence consisted of counting the cases on one side or the other was the calculus of games of chance applicable. Bernoulli had hopes of expanding that domain, but frankly acknowledged that such

situations were rare. In all other instances, of everyday reasoning, what was entailed was a scale of evidence whose lower bound was a complete absence of evidence in favor of a proposition. Both Bernoulli and Lambert, however, proceeded as though both types of evidence could be assessed by counting the cases; the only material distinction was that, for pure arguments, the cases on one side were mute. Their focus on counting cases, however, confused both them and their readers. Note Hacking's (1974) reference to "the 10% of remaining cases" in the pure argument about the newspaper being of today's date. What "cases"? Lambert's treatment was abstract; he was even stingier with examples than Bernoulli. But it is not clear that there exist pure arguments which can be evaluated by counting cases.

Bernoulli's concept of pure arguments could be characterized as formally adequate in principle: adequate to the task of representing ignorance, by virtue of non-additivity (allowing for weights of evidence less than exhaustive); I say "formally" because there is no indication that such arguments are susceptible to numerical measurement, and "in principle" because nonadditivity contradicts Bernoulli's definition of probability as a proportion and because his rule for combining evidence was inappropriate. The truth to be salvaged from the concept of a pure argument is that formalizing reasoning in "civil, moral, and economic matters" would require a scale from no evidence, or total ignorance, at one end to conclusive evidence, or certainty, at the other, not a scale with certainty at both ends. But there is no reason to believe that any useful quantification of such a scale is possible. Both Bernoulli and Lambert failed to recognize—like almost everyone after them—that two distinct scales are involved. The numbers from the pure and mixed cases mean different things. No meaning can be ascribed to results from a formula which adds inches and pounds. So far as I can tell, Edith Sylla is the only one to have acknowledged even the possibility of the problem:

Perhaps more fundamental than the particular formula for combining different sorts of evidence is the question whether the weight of different sorts of evidence can be combined reasonably at all. In proposing to calculate mathematically what decisions should be made in political, moral, and economic life, Bernoulli assumes implicitly that he can measure the weight of all his evidence on a single scale in any given problem. In this example of the stabbing, is it really clear that evidence of cloak color can be weighed on the same scale as evidence of behavior under questioning? (1990, pp. 31-32)

3.6 Implications for Future Developments

The concept of nonadditive probabilities, though entirely unremarkable to Bernoulli and Lambert, was virtually stillborn. Shafer (1978) believes that the last authors even to notice the possibility of nonadditive probability were Prevost and Lhuilier (1797). They praised Lambert for his improvement of Bernoulli's formulas but did not adopt his nonadditive rules; in fact, they took a dim view of the whole project of applying the calculus of probabilities to testimony. Shafer also observes that even

Todhunter (1865), in his encyclopedic history, omitted any reference to nonadditive probability in his discussion of Lambert's work.

Bernoulli's contribution to the concept of probability was thus deeply ironic. He was apparently the first writer since Arnauld (1662) to use *probability* in aleatory contexts, and this time the linking of chance and knowledge caught on. He was also the first to attempt a formal calculus for epistemic probabilities that went beyond the crude three-point scale proposed by Leibniz, but his rules for combining epistemic and aleatory probabilities were neither formally adequate nor practically useful. The subsequent neglect of what he called pure arguments meant that his successors implicitly assumed both kinds of probabilities amenable to the calculus of gambling. Thus Bernoulli's use of the same word for the two concepts historically overpowered his own efforts to keep them separate.

The subsequent history of the theory of probability cannot be laid to a mere issue of labeling, however. Bernoulli's linking of probability and frequency through his law of large numbers clinched the disappearance of nonadditive probability: If they can be interpreted as relative frequencies, then all probabilities are constrained to a 0–1 scale where the lower bound represents impossibility or the certainty of the contrary.

Besides the law of large numbers, several other factors contributed to the extinction of the concept of nonadditivity. Most obviously, it simply did not square with the established calculus of chances, which could already boast a rapidly developing algebra. The nonadditive probabilities of Bernoulli and Lambert lacked correspondingly convenient or compelling mathematics; it was never clear how practical problems might be solved with them. If Bernoulli, or someone soon after him, had succeeded in developing not only a viable epistemic calculus but a scale to provide unambiguous numbers, the subsequent history might have been much different. There might never indeed have been any concept of statistical inference: for in place of one dualistic concept, we might have ended up with two distinct concepts, which would eventually have pulled for separate labels.

What Shafer (1978) characterizes as the subtlety of Bernoulli's thought might less charitably be regarded as confusion. But Bernoulli's broad philosophical perspective led him at least to consider pure arguments and nonadditive probabilities. This aspect of his thought was lost on his successors, Shafer suggests, largely because the next two illustrious writers on the subject, Pierre de Montmort and Abraham de Moivre, were mathematicians much more than philosophers. Montmort was working on his *Essay d'Analyse sur les Jeux de Hazard* before *Ars Conjectandi* was published; he knew of it only through eulogies published in France after Bernoulli's death in 1705. These reviews understandably gave the impression that Bernoulli had succeeded in applying the calculus of chance to practical affairs. Montmort in his preface was frankly skeptical about application of mathematical probability to civil matters and didn't treat them in his book, though he professed confidence, based on Bernoulli's reputation, that the latter had done a perfect job of it.

De Moivre's principal contribution in the present context was to establish *probability* as the fundamental term of the theory. His first treatise on the subject, *De*

Mensura Sortis, published in 1711, was inspired by the first edition of Montmort's book 3 years earlier and, like that book, spoke mainly of expectation. Five years after the publication of *Ars Conjectandi*, however, de Moivre published *The Doctrine of Chances*, the first modern-sounding textbook of probability. He defined probability directly as a ratio, in fact as a proportion, which insures additivity. Shafer (1978) notices that this is the first statement of additivity as a rule. De Moivre's text, like Montmort's, was mathematical, rather than a wide-ranging epistemological treatise like Bernoulli's. Its success as a textbook of probability, however, helped very much, Shafer argues, to establish probability, and additive probability, as the fundamental concept. For these reasons, Shafer suggests, the disappearance of Bernoulli's broader numerical conception could almost be attributed to the delay in publication of *Ars Conjectandi*. Had Montmort and de Moivre known just how far Bernoulli was from succeeding in his ambition to make the calculus of chances applicable to the uncertainties of everyday life, their own claims for the potential of the theory would presumably have been much more circumspect. Once it was announced, on such authority, that that extraordinary goal was within reach, there was little chance of turning back to reexamine the premises.

The promise contained in Bernoulli's simple act of labeling was enormous. All nondemonstrative knowledge—which was already coming to mean all knowledge except mathematics (and Scripture)—was (merely) probable; but now this evidently meant that the uncertainty of all our inferences was amenable to precise calculation—just as Cowles (1989) believes (see Chap. 1). With the formulation of this dualistic concept, the vague hopes, described in the previous chapter, for the mathematical formalization of all reasoning stood suddenly to be realized, through the concept of statistical inference. Laplace, who made the principally influential contribution to implementing it, did not even exclude mathematics from its domain:

Almost all our knowledge is but probable, and in the small number of things we can know with certainty, in the mathematical sciences themselves, the principal means of arriving at truth, induction and analogy, are based on probabilities; as a result the entire system of human knowledge is rooted in [that] theory. (1816, pp. 1-2)

Bernoulli's two separate concepts could not have promised so much. What was required was a single concept—a single word, at least—that embraced both the epistemic meaning and the additive calculus of chances. To make the integration of these two aspects plausible was a conceptual task which virtually everyone undertook—there was little obvious payoff in resisting it—but it was not a task that was accomplished either so easily or quickly as it might appear. (The corresponding struggle in ontogenesis is examined briefly in Chap. 10.) In general, wherever a mathematical concept of probability was clearly meant, additivity was invariably assumed, from the late eighteenth century on. Ancillon (1794) wrote as though mathematical probability was necessarily based on an enumeration of equally likely cases and thus questioned any application outside games of chance. His point was on target: We have long since given up attempting Laplacean calculations like that, for example, of entering another ice age within the next century; and other

applications, such as the use of urn models in epidemiology or thermodynamics, indeed take games as analogs.

Jonathan Cohen (1980) has found evidence in the epistemological and legal literature of the seventeenth through the nineteenth centuries for the persistence of a concept that is implicitly nonadditive (or non-Pascalian, in his terminology). His examples consist of references to a scale of probability which ranges from certainty or provability or adequate evidence for proof downward to nonprovability or no evidence, rather than to disprovability. It is difficult to tell whether the writers he refers to were implicitly thinking of the old, nonmathematical meaning of probability or whether they simply failed to bring the issue sharply enough into focus to see the discrepancy between the scales. A single example may suffice.

Francis Bacon was writing before the advent of the calculus of gambling, and his references to probability can best be understood as non-Pascalian. Mill, however, adding the principle of additivity in his interpretation of Baconian ideas, made them look more foolish. In analogical reasoning, Bacon would have agreed that a resemblance between A and B on one property gave some probability to the proposition that they were alike in others:

But Mill went on to claim that, if after much observation of B, we find that it agrees with A in nine out of ten of its known properties, we may conclude with a probability of nine to one, that it will possess any given derivative property of A. Rather naively he assumed not only that all the properties involved would be of equal significance but also that they would be additive, so that he would be entitled to measure probability here by an arithmetical ratio. In fact both assumptions are highly precarious. The various factors relevant to testing a particular causal hypothesis may differ very much from one another in importance, as Bacon acknowledged in his discussion of prerogative instances. And the Baconian probability of a generalization, since it is determined by the generalization's capacity to resist falsification under possible combinations of relevant causal factors, is inherently non-additive. Two causal factors may have an explosive (or mutually neutralizing) effect in combination which neither even remotely approaches in isolation. (Cohen, 1980, p. 227)

The simultaneous presumption of additivity and of general epistemic application thus created a tension within the concept of probability from the start. The history of the concept up to the present can be understood largely as a struggle with that tension. In the first two centuries it would take the form of a gradual withering of interest in the traditional epistemic applications. Criticisms like those of d'Alembert (1767) or Ancillon (1794) would no longer seem relevant. Todhunter (1865) dismissed Ancillon's memoir as not worth discussing, and hardly anyone except Keynes (1921/ 1973) appears to have consulted it since. And Keynes' bemused appraisal of d'Alembert—"His opposition to the received opinions was, perhaps, more splendid than discriminating" (p. 90n)—may have been as charitable as any he received (at least until recently; see Swijtink, 1986).

The decline of the original concept can be traced in the probability of testimony and legal judgments. Testimony was a natural—indeed paradigmatic—application of probability in its original meaning, as the reliability of human authority, and most writers up to the twentieth century had something to say about it. It is probably also the most arbitrary, fruitless chapter in the history of the theory. The additive model posed problems both for the measurement of probabilities—in this case, the

credibility of witnesses—and in rules for their combination. The credibility of a witness is not logically constrained to conform to a scale representing proportions, where testimony for and against a proposition must sum to 1, regardless of how well informed the witness is. On such a model, as was remarked above, the testimony of one who is totally ignorant carries as much “weight” as an expert’s. It was usually also unclear, as Keynes (1921/ 1973) pointed out, whether a credibility coefficient represented the probability that if A is true, X would assert it, or the probability that, if X asserts A , it is true. Regarding the combination of probabilities, it was generally assumed that if two witnesses were independent in the sense of there being no collusion between them, the probability that both spoke the truth was the product of the separate reliabilities. As Keynes noted, however, this assumption ignores the relation between the propositions asserted by the two witnesses—whether they are contradictory, or the same statement, or mutually irrelevant; in fact, it turns out that the product rule holds only for propositions a priori equally likely to be true as false (Keynes, 1921/ 1973, p. 182).

The most notorious contribution to the probability of testimony was made by John T. Craig (1699/ 2011), who claimed that his calculation “consolidates the foundations of Christianity and, at the same time, completely destroys Judaism” (1699/ 2011, p. 6). Todhunter (1865, p. 54) cites the anonymous *Treatise on Probability* by Lubbock and Drinkwater,²⁴ published in the Library of Useful Knowledge around 1830:

“It is not necessary to do more than mention an essay, by Craig, on the probability of testimony, which appeared in 1699, under the title of ‘*Theologiae Christianæ Principia Mathematica* [Mathematical Principles of Christian Theology].’ This attempt to introduce mathematical language and reasoning has the appearance of an insane parody of Newton’s *Principia*, which then engrossed the attention of the mathematical world. The author begins by stating that he considers the mind as a movable, and arguments as so many moving forces, by which a certain velocity of suspicion is produced, &c. He proves gravely, that suspicions of any history, transmitted through given time (*cæteris paribus*), vary in the duplicate ratio of the times taken from the beginning of the history, with much more of the same kind with respect to the estimation of equable pleasure, uniformly accelerated pleasure, pleasure varying as any power of the time, &c &c.”

Todhunter goes on: “It seems that Craig concluded that faith in the Gospel so far as it depended on oral tradition expired about the year 800, and that so far as it depended on written tradition it would expire in the year 3150” (pp. 54–55). Laplace (1816, pp. 139–141) took on Craig’s argument and turned it on its head. He accepted Craig’s premise that the credibility of Biblical tales diminished with each successive retelling; but whereas Craig had argued that the infinite payoffs for believing them kept their credibility finite, Laplace argued that the promise of infinite reward made the temptation to lie approach infinity and hence that the testimony was worthless.

There was a time when applications of the probability of testimony and judgments were taken very seriously. Condorcet was imprisoned because his work on

²⁴The authors are sometimes listed (e.g., in Walker, 1929) as Lubbock and Bethune, but the second author’s name was actually Drinkwater-Bethune.

the probability of jury decisions was somehow seen as counterrevolutionary, and he poisoned himself to avoid the guillotine. Laplace avoided writing on the probability of testimony between 1783 and 1812 evidently because of what had happened to Condorcet (K. Pearson, 1978). Schneider (1987) notes that after Poisson (1837), however, it was only textbook writers and not scientists who paid any attention to the probability of testimony. Aleatory probabilities were soon to become important in the physical and social sciences, and the probability of testimony had nowhere to go.

Daston (1988) argues at some length that the demise of classical probability was assisted, in a paradoxical way, by the associationist psychology of the empiricists. For the British empiricists, especially Hume and Hartley, the strength of a reasonable man's beliefs was a reflection of the frequency with which a particular association had presented itself to him. Subjective probability was thus a mirror of objective probability and was hardly distinguishable from it. Probability theory was also as much descriptive as prescriptive, though these authors acknowledged the need for occasional corrections in the process. The French Idéologues, on the other hand, perhaps more cynical under the more direct influence of Descartes, were more impressed by the distortions of belief from childhood experience, prejudice, and the like. Their pessimism led initially to frantic lobbying for general education in probability theory, so that all—especially, as far as Condorcet was concerned, “children and the people”—might reason as the elite of enlightened men. But Daston argues that the increasing emphasis on error and distortion undermined the claim of probability theory to be descriptive of human thought and drove a wedge between the subjective and objective aspects of probability. Daston's summary paraphrases a remark of Leonard Peikoff's in a somewhat different context: “What the mind giveth, the mind taketh away” (Peikoff, 1979, p. 162). Application of probability theory to tribunals stretched the system beyond the limit of plausibility. After Poisson, applications of probability in civil and moral matters shifted from the credibility of individuals to the stability of aggregates, in Quetelet's work on social statistics (Chap. 5).

As epistemic applications like testimony declined in importance and were relegated to textbook discussions, and aleatory probabilities were becoming more important in scientific applications, a new terminology arose to mark the two different contexts of use. Poisson (1837) first proposed the term *probability* for the epistemic meaning and *chance* for aleatory probability. “Probability depends on the knowledge we have of an event; it may be different for the same event and different persons” (p. 30); somewhat like Jeffreys (1961), he used the word *chance* to refer to “events in themselves, independent of our knowledge of them” (p. 30).

The distinction was developed much more fully and explicitly by Cournot (1843, 1851/1956), who blamed Bernoulli for starting the confusion by his use of subjective terms like *conjecture* in the context of mathematical probability. Cournot insisted that subjective and objective probability must be kept distinct, but he did not devalue the former; indeed, he said, all nondeductive reasoning makes use of it. Subjective probability arises from imperfections in our knowledge; it cannot be called objective because the particular balance of evidence known to one person

will not necessarily correspond to the ratio of favorable and unfavorable conditions in reality, nor to the balance of evidence known to another individual. Cournot (1851/ 1956) argued, like Ancillon (1794), that subjective, or philosophical, probability cannot be quantified for the reason that all possible laws and conditions cannot generally be enumerated in such a way as to permit calculation of precise probabilities, at least without a considerable degree of arbitrariness. To objective probability, on the other hand, he gave a frequency definition. He referred to it as “physically impossible” for the long-run relative frequency to diverge by a finite amount from the *a priori* value of p —thus according Bernoulli’s Theorem empirical status—and the conditions for objective or mathematical probability were said to hold independently of our knowledge. By Cournot’s time the duality in the concept of probability was thus explicit and well established. In spite of the separate designations of “subjective” versus “objective” or “probability” versus “chance,” the concept was truly a dual one, for in critical applications aleatory probability was taken as a measure of epistemic probability.

A split in concepts and, later, theories of probability was a predictable consequence of Bernoulli’s attempt to unite these two aspects in one concept. If the mathematics of probability was henceforth to be exclusively that of games of chance, the meaning of probability was still to remain fundamentally epistemic—at least until a point, a century or two later, when the aleatory aspect had been so thoroughly assimilated that it was possible for philosophers taking only a contemporary view to wonder indeed which aspect was more fundamental. With the emergence of the dualistic concept of probability, then—an aleatory mathematics grafted onto an epistemic meaning—there arose the possibility of rival theories seeking to explicate the concept. Some—a minority—would claim, with Bernoulli and his generation, that it was fundamentally an epistemic concept, with the calculus of gambling as a special application. Under the influence of positivism in the late nineteenth century, the prevailing view would come to be that the frequency meaning was fundamental, and that popular epistemic connotations were to be jettisoned from science as metaphysical or psychological. In the twentieth century, of course, some would argue that the meaning didn’t matter, as long as we could do the calculations.

In the meantime Bernoulli’s law of large numbers would prompt a search for a way to invert it—to get *a priori* probabilities out of observed frequencies. This problem is the first example of statistical inference. Solutions to it would prove to depend crucially on the definition of probability that was adopted. In the eighteenth century, when the first solutions were proposed by Bayes and Laplace, it was inevitable that they be based on a Bernoullian view of probability as epistemic. Their approach would be an example of what would be called, in the twentieth century, Bayesian inference. That much right to the title it has; but, because their work needs to be distinguished from twentieth-century Bayesian statistics, and also because Laplace played a major role in its development, I refer to their contributions jointly by the nonstandard label of the classical theory of statistical inference. The next chapter will show how the concept of statistical inference, exploiting the ambiguity of the dualistic concept, appeared to solve the skeptical problem of induction that had been created by the externalizing transformation of knowledge in the preceding century.

References

- Allport, G. W. (1962). The general and the unique in psychological science. *Journal of Personality*, 30, 405–422.
- Ancillon (J. P. F.). (1794). Doutes sur les bases du calcul des probabilités [doubts on the bases of the calculus of probabilities]. *Mémoires de l'Académie Royale des Sciences et Belles-lettres de Berlin*, 3, 3–32.
- Anonymous (1706). Éloge de M. Bernoulli, cy-devant Professeur de Mathématique à Bâle [Eulogy of Mr. Bernoulli, formerly Professor of Mathematics at Basel]. *Journal des Scavans*, 34 (Part I), 126–139.
- Arnauld, A. (1662). *La logique, ou l'art de penser* [Logic, or the art of thinking]. Paris: Jean de Launay.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Bernoulli, J. (1713). *Ars conjectandi* [The art of conjecturing]. Basel: Thurneysen.
- Blackburn, S. (1980). Review of The Probable and the Provable. *Synthèse*, 44, 149–159.
- Brakel, J. v. (1976). Some remarks on the prehistory of the concept of statistical probability. *Archive for History of Exact Sciences*, 16, 119–136.
- Brown, R. (1987). History versus Hacking on probability. *History of European Ideas*, 8, 655–673.
- Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon Press.
- Cohen, L. J. (1980). Some historical remarks on the Baconian conception of probability. *Journal of the History of Ideas*, 41, 219–231.
- Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités* [Exposition of the theory of chances and probabilities]. Paris: Hachette.
- Cournot, A. A. (1956). *An essay on the foundations of our knowledge*. New York: Liberal Arts Press. (Original work published 1851).
- Cowles, M. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, NJ: Erlbaum.
- Craig, J. (2011, June). Principes mathématiques de théologie Chrétienne [Mathematical principles of Christian theology] (J.-Y. Guillaumin, Trans.). *Journ@l Électronique d'Histoire des Probabilités et de la Statistique*, 7(1). (Original work published 1699).
- d'Alembert, J. le R. (1767). Doutes et questions sur le calcul des probabilités [Doubts and questions on the calculus of probabilities]. In *Mélanges de littérature, d'histoire, et de philosophie* (Vol. 5, pp. 275–304). Amsterdam: Chatelain.
- Daston, L. J. (1983). Mathematical probability and the reasonable man of the eighteenth century. In *History and philosophy of science: Selected papers* (pp. 57–72). New York: Annals of the New York Academy of Science.
- Daston, L. J. (1987). The domestication of risk: Mathematical probability and insurance 1650–1830. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 237–260). Cambridge, MA: MIT Press.
- Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton: Princeton University Press.
- Daston, L. J. (1992). The doctrine of chances without chance: Determinism, mathematical probability, and quantification in the seventeenth century. In E. N. Hiebert, M. J. Nye, J. L. Richards, & R. H. Stuewer (Eds.), *The invention of physical science: Intersections of mathematics, theology, and natural philosophy since the seventeenth century: Essays in honor of Erwin N. Hiebert* (pp. 27–50). Dordrecht: Kluwer Academic.
- David, F. N. (1955). Studies in the history of probability and statistics. I. Dicing and gaming (A note on the history of probability). *Biometrika*, 42, 1–15.
- David, F. N. (1962). *Games, gods and gambling*. New York: Hafner.
- Deman, T. (1933). Probabilis [Probable]. *Revue des Sciences Philosophiques et Théologiques*, 22, 260–290.
- de Moivre, A. (1718). *The doctrine of chances*. London, UK: W. Pearson.
- Dutka, J. (1988). On the St. Petersburg paradox. *Archive for History of Exact Sciences*, 39, 13–39.
- Foucault, M. (1973). *The order of things*. New York: Random House.

- Franklin, J. (2001). *The science of conjecture: Evidence and probability before Pascal*. Baltimore: Johns Hopkins University Press.
- Freudenthal, H. (1980). Huygens' foundations of probability. *Historia Mathematica*, 7, 113–117.
- Garber, D., & Zabell, S. (1979). On the emergence of probability. *Archive for History of Exact Sciences*, 21, 33–53.
- Grip, J. (2009). *Knightian Uncertainty*. Master's thesis, Uppsala University Institute for Information Science.
- Hacking, I. (1974). Combined evidence. In S. Stenlund (Ed.), *Logical theory and semantic analysis* (pp. 21–37). Dordrecht: Reidel.
- Hacking, I. (1975). *The emergence of probability*. Cambridge, UK: Cambridge University Press.
- Hald, A. (1990). *A history of probability and statistics and their applications before 1750*. New York: Wiley.
- Holt, R. R. (1962). Individuality and generalization in the psychology of personality. *Journal of Personality*, 30, 377–404.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon Press. (1st ed., 1939).
- Jorland, G. (1987). The Saint Petersburg paradox. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 157–190). Cambridge, MA: MIT Press.
- Keynes, J. M. (1973). *A treatise on probability*. New York: St. Martins Press. (Original work published 1921).
- Lambert, J.-H. (2002). *Nouvel organon: Phénoménologie* [The new organon: Phenomenology] (G. Fanfalone, Trans.). Paris, France: Vrin. (Original work published 1764).
- Laplace, M. le Comte de (1816). *Essai philosophique sur les probabilités* [Philosophical essay on probabilities] (3rd ed.). Paris, France: Courcier.
- Loosli-Usteri, M. (1931). La conscience du hasard chez l'enfant [The child's understanding of chance]. *Archives de Psychologie*, 23, 45–66.
- McKeon, R. (Ed.). (1941). *The basic works of Aristotle*. New York: Random House.
- Monahan, J., & Steadman, H. J. (1996). Violent storms and violent people: How meteorology can inform risk communication in mental health law. *American Psychologist*, 51, 931–938.
- Montmort, P. R. de (1713). *Essay d'analyse sur les jeux de hazard* [Analytical essay on games of chance] (2nd ed.). Paris, France: Quillau.
- Murphy, A. H. (1991). Probabilities, odds, and forecasts of rare events. *Weather and Forecasting*, 6, 302–307.
- Partridge, E. (1966). *Origins* (4th ed.). New York: Macmillan.
- Pearson, K. (1978). *The history of statistics in the 17th and 18th centuries, against the changing background of intellectual, scientific, and religious thought* (E. S. Pearson, Ed.). London, UK: Griffin.
- Peikoff, L. (1979). The analytic-synthetic dichotomy. In A. Rand (Ed.), *Introduction to objectivist epistemology* (pp. 117–164). New York: New American Library. (Original work published 1967).
- Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités* [Investigations on the probability of judgments in criminal and civil matters, preceded by general rules of the calculus of probabilities]. Paris, France: Bachelier.
- Porter, T. M. (1986). *The rise of statistical thinking 1820–1900*. Princeton: Princeton University Press.
- Prévost, P., & Lhuillier, S. A. S. (1797). Mémoire sur l'application du calcul des probabilités à la valeur du témoignage [Memoir on the application of the calculus of probabilities to the value of testimony]. *Mémoires de l'Académie Royale des Sciences et Belles Lettres de Berlin*, 6, 120–152.
- Pye, M. (2015). *The edge of the world: A cultural history of the North Sea and the transformation of Europe*. New York: Pegasus Books.
- Sambursky, S. (1956). On the possible and the probable in ancient Greece. *Osiris*, 12, 35–48.

- Schneider, I. (1980). Why do we find the origin of a calculus of probability in the seventeenth century? In J. Hintikka, D. Gruender, & E. Agazzi (Eds.), *Probabilistic thinking, thermodynamics and the interaction of the history and philosophy of science. Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science, Vol. 2* (pp. 3–24). Dordrecht: Reidel.
- Schneider, I. (1987). Laplace and thereafter: The status of probability calculus in the nineteenth century. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 191–214). Cambridge, MA: MIT Press.
- Schneider, I. (1988). The market place and games of chance in the 15th and 16th centuries. In C. Hay (Ed.), *Mathematics from manuscript to print, 1300–1600* (pp. 220–235). Oxford: Clarendon Press.
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.
- Shackle, G. L. S. (1961). *Decision, order and time in human affairs*. Cambridge, UK: Cambridge University Press.
- Shafer, G. (1976a). *A mathematical theory of evidence*. Princeton: Princeton University Press.
- Shafer, G. (1976b). Review of *The emergence of probability*. *Journal of the American Statistical Association*, 71, 519–521.
- Shafer, G. (1978). Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*, 19, 309–370.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24, 449–474.
- Shapiro, B. J. (1983). *Probability and certainty in seventeenth-century England*. Princeton, NJ: Princeton University Press.
- Shuman, F. G. (1989). History of numerical weather prediction at the National Meteorological Center. *Weather and Forecasting*, 4, 286–296.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
- Swijtink, Z. G. (1986). D'Alembert and the maturity of chances. *Studies in History and Philosophy of Science*, 17, 327–349.
- Sylla, E. (1990). Political, moral, and economic decisions and the origins of the mathematical theory of probability: The case of Jacob Bernoulli's *The Art of Conjecturing*. In G. M. von Furstenberg (Ed.), *Acting under uncertainty: Multidisciplinary conceptions* (pp. 19–44). Boston: Kluwer Academic.
- Sylla, E. D. (1998). The emergence of mathematical probability from the perspective of the Leibniz-Jacob Bernoulli correspondence. *Perspectives on Science*, 6(1–2), 41–76.
- Sylla, E. D. (2003). Business ethics, commercial mathematics, and the origins of mathematical probability. In M. Schabas & N. de Marchi (Eds.), *Oeconomies in the age of Newton* (pp. 309–337). Durham, NC: Duke University Press.
- Sylla, E. D. (2006). Introduction and notes to *The Art of Conjecturing*. Baltimore: Johns Hopkins University Press.
- Todhunter, I. (1865). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge, UK: Macmillan.
- Venn, J. (1888). *The logic of chance* (3rd ed.). London, UK: Macmillan. (1st ed., 1866).
- Walker, H. M. (1929). *Studies in the history of statistical method*. Baltimore: Williams and Wilkins.
- Weaver, W. (1963). *Lady luck: The theory of probability*. Garden City, NY: Doubleday Anchor.
- Weisberg, H. I. (2014). *Willful ignorance: The mismeasure of uncertainty*. New York: Wiley.
- Yushkevich, A. P. (1987). Nicholas Bernoulli and the publication of James Bernoulli's *Ars Conjectandi. Theory of Probability and its Applications*, 31, 286–303.

Chapter 4

The Classical Theory of Statistical Inference



In applying the traditional, epistemic concept of probability to games of chance, Bernoulli opened up an area of huge promise, stopped only by an apparently technical problem. The promise, it is true, is not one he would necessarily have endorsed, but to those following him his work clearly held out the prospect of a mathematical formalization of all knowledge and inference, based in some way on the calculus of gambling. Extension of the gambling model to even the most straightforward application, however, confronted the problem of identifying the equally likely cases corresponding to the alternative outcomes with dice or coins. Hudde's scheme of allotting equal chances of death in 6-month intervals across the lifespan happened to achieve a plausible fit to available data, but it lacked any real logical basis. Bernoulli himself drew some hope in the notable case where extensive data were collected, namely mortality statistics. His intuitive idea, clearly, was that the relative frequency with which something happened ought to provide an estimate of its probability of occurrence. What he succeeded in proving was tantalizingly close: that, by sufficiently extending the number of trials, we can, with a probability arbitrarily close to 1, make the relative frequency of occurrence of the event arbitrarily close to the actual probability p . Alternatively, if a fixed number of trials are to be made, the theorem can be used to calculate probabilistic limits for the relative frequency that will result.

The frustration of Bernoulli's Theorem, however, is that these limits do not translate into limits on the value of p given some sequence of trials. And in real-life applications, such as insurance, it is the frequency which is given, and the probability of an individual event which is sought. In the language of the day, what was sought were the probabilities of various underlying causes, given the observed event.¹ What was more readily calculable, on the other hand, were the probabilities of various events or outcomes, given the "cause" (e.g., a fair die). It is true, as was

¹Felix Mendelssohn's grandfather Moses has been credited by at least one authority with priority in linking probability with cause and effect (see Fagot, 1980).

indicated in Chap. 2, that this way of speaking was already being undermined. But language proceeded, as it were, ahead of thought: Although from this point it grew increasingly unacceptable to *speak* of causes in the old sense of agency, still hardly anyone *thinks* as a Humean. In other respects, there was, philosophically, no special difficulty with the locution of the probability of a cause: Probability was still a thoroughly epistemic concept, and causes were valid and important objects of uncertainty. Mathematically, on the other hand, the notion offered a difficulty: If the additive calculus of chances were to be applied to the probability of causes, then it was entailed that causes have a probability distribution. Though it wasn't framed in exactly these terms at the time, it was necessary that the causes themselves be the outcome of some random process, before the aleatory calculus could be brought to bear on them. A significant logical leap was thus required for the solution, and it is not so surprising that it waited another half century after Bernoulli.

The solution was offered independently,² about a decade apart, by Bayes and Laplace. Their contributions constitute the first examples of statistical inference. If the problem was that there was no chance mechanism which determined the a priori value of p , one solution was just to imagine that there was. This was essentially Bayes' approach, to frame the situation as a two-stage experiment, in which the first stage yielded the value of p governing the subsequent trials. Laplace's solution, though it can be reinterpreted along similar lines as a two-stage experiment, actually relied more directly on sheer assumption. Each scheme had its advantage—Bayes' analytical, Laplace's rhetorical. Curiously, in view of the familiarity of his name in the late twentieth century, Bayes' paper was very little noticed for a century and a half; it was Laplace's model which captured the imagination of the nineteenth century.

4.1 Bayes

Curiously also, little is known about Thomas Bayes himself, a Nonconformist³ minister (Dale, 1991 and Bellhouse, 2004, have compiled the available biographic minutiae). His lack of either mathematical or theological publications has occasioned

²That is at least a reasonable presumption. Laplace does make passing reference to Bayes in later work, but his approach is so different from Bayes, and Bayes' paper was so little noticed by others, that there is no reason to assume he had seen it before reaching his own solution.

Stigler (1983) has uncovered some evidence to suggest that the original author of Bayes' Theorem—of statistical inference, actually—might have been the mathematician Nicholas Saunderson. The remarkable Saunderson, who at age 30 succeeded to the chair that Newton had once held, had been blind since the age of 1.

³The issue was the *Book of Common Prayer*, compiled by Thomas Cranmer, Archbishop of Canterbury, in 1549 for use by the Church of England. The book passed in and out of favor with changing regimes; an Act of Uniformity in 1662 requiring its use was resisted by some 2000 clergy, who were forced from their positions. An Act of Tolerance in 1689 allowed Dissenting or Nonconformist congregations to meet, provided that the chapels were licensed (Bellhouse, 2004).

some question as to how he could have been elected to the Royal Society in 1742; he is, however, now credited with an anonymous 1736 paper defending Newton's method of fluxions (calculus) against Berkeley's attacks on it as metaphysical (Dale, 1991). His "An Essay Towards Solving a Problem in the Doctrine of Chances," published 2 years after his death, was read to the Society by Richard Price, who, ironically, is little known today but was a hero in his time. An ardent pamphleteer for liberty, he was responsible for coining the name "United States of America" (Hacking, 1990). He declined the invitation in 1778 by the American Congress to become a citizen and to help with administering the finances of the Revolutionary War, but he was one of two men to accept an honorary LL.D. from Yale in 1781, the other being General George Washington (Pearson, 1978). Price edited Bayes' paper for publication, replacing Bayes' introduction with his own and adding an appendix, so it is not always clear which ideas Bayes would have supported.

Bayes approached his subject in the classical manner, with the first part of his paper comprising formal definitions and theorems of elementary probability theory. These include the basic formula for conditional probability: "If of two subsequent events [Bayes did not explain this peculiar locution] the probability of the 1st be a/N , and the probability of both together be P/N , then the probability of the 2nd on the supposition the 1st happens is P/a " (Bayes, 1763, p. 379). Bayes' probabilities here are ratios, since he defined probability as a ratio of expectations (see Chap. 3); in modern terms, if A and B are two events, we say

$$P(B|A) = P(A \& B) / P(A).$$

We would also nowadays unhesitatingly write

$$P(A|B) = P(A \& B) / P(B).$$

as formally similar. If A is antecedent to B , however, Bayes thought the latter proposition required separate proof. In his formulation,

If there be two subsequent events, the probability of the 2nd b/N and the probability of both together P/N , and it being discovered that the 2nd event has happened, from hence I guess that the 1st event has also happened, the probability I am in the right is P/b . (p. 381)

Shafer (1982) has subjected Bayes' reasoning on this point to a close analysis, tentatively concluding that the time dimension does not serve to distinguish these propositions after all. In any event, both Bayes and we accept both propositions as true. Then clearly

$$P(A|B)P(B) = P(B|A)P(A).$$

We can see the relevance to Bayes' ultimate problem if we rewrite this expression in more suggestive terms. Let D (for data) represent the aggregate outcome on a series of binomial trials, and H denote hypotheses about the value of p . Suppose we distinguish two such hypotheses: H_0 , that $p \leq 0.5$; H_1 , that $p > 0.5$. Logically, one or the other of these has to be true, so we can write

$$P(D) = P(D \& H_0) + P(D \& H_1) = P(D|H_0)P(H_0) + P(D|H_1)P(H_1).$$

And since

$$P(H_0|D)P(D) = P(D|H_0)P(H_0),$$

we have

$$P(H_0|D) = \frac{P(D|H_0)p(H_0)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)}.$$

If the range of alternatives partitioned into more than two discrete hypotheses, we could write

$$P(H_0|D) = \frac{P(D|H_0)p(H_0)}{\sum P(D|H_i)p(H_i)}.$$

More often, and particularly in Bayes' problem, we might think of the continuum from 0 to 1 as representing an infinity of hypotheses about the value of p , in which case the probabilities would be replaced by probability densities and the summation by integration:

$$f(p_0|D) = \frac{f(D|p_0)f(p_0)}{\int_0^1 f(D|p)f(p)dp}.$$

The denominator, a closed integral, is essentially a constant of proportionality.

These various formulas are examples of what is now known as Bayes' Theorem. The theorem is useful in general when we have $P(A|B)$ and wish to find $P(B|A)$. In the present, typical case, what is readily available is $P(D|H)$, the probability of various outcomes assuming some particular hypothesis about p to be true; and what is wanted are statements about $P(H|D)$, the probability of various values of p given some observed outcome. The theorem as such does not actually appear in Bayes' paper, since he argued in a geometric rather than an analytic mode. His development is too complex to present here, and the above formulas, due to Laplace, are more familiar to modern readers.⁴

Bayes framed the basic problem as follows: “*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of

⁴ Since Bayes was obviously facile with integral calculus, it is curious that he resorted to so much more cumbersome a mode of argument. Dale (1991) quotes the following (“faintly chauvinistic”) explanation from Timerding, who was writing in German in 1908: “To understand Bayes’ presentation, one must remember that in England the integral notation was taboo, since its author Leibniz amounted to a plagiarist of Newton” (my translation from Dale, 1991, p. 21).

probability that can be named" (1763, p. 376). His chance mechanism for generation of the value of p was a table representing the unit square, onto which a ball W is tossed (Fig. 4.1). A line os is drawn through the point where the ball comes to rest, parallel to one side, AD , of the square. Then a second ball, O , is tossed onto the table, and Bayes defined the event M as having occurred if O comes to rest between os and AD . If the distance between os and AD is p , then p is the probability of the event M . He then supposed that in n trials of tossing O on the table, we have observed r occurrences of M , without knowing the results of the first toss; and he sought the probability that, on the basis of these observations, p lay in any given interval (a, b) .

By the binomial formula, the probability of r occurrences of M in n trials is

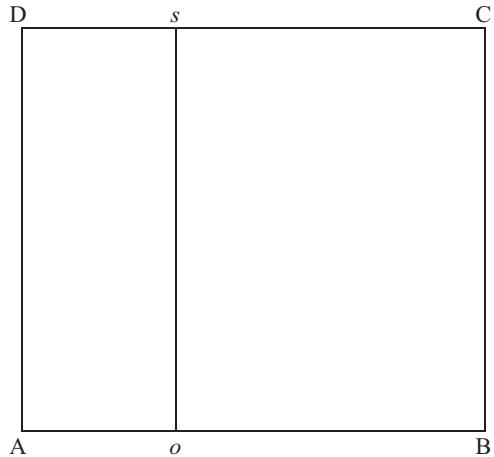
$$\frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}.$$

The probability that p lies in some "small" interval dp is just dp , since os is equally likely to fall anywhere on the unit line. Hence the compound probability that M occurs r times in n trials and that p lies in dp is

$$\frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} dp.$$

The conditional probability that p lies in (a, b) , given r occurrences of M , is then, by application of Bayes' Theorem,

Fig. 4.1 Bayes' "billiard table"



$$\frac{\int_a^b p^r (1-p)^{n-r} dp}{\int_0^1 p^r (1-p)^{n-r} dp},$$

the binomial coefficients canceling. This exact expression again is not Bayes' own, but he arrived at the same quantity represented by areas.

Bayes' reasoning was rather careful,⁵ and impressive—"d'une manière fine et très ingénieuse, quoiqu'un peu embarrassée," Laplace (1816, p. 215) himself would say; for the situation he set up, his solution is correct. For the inversion of Bernoulli's Theorem in general, however, it constitutes but an argument from analogy. Bayes made this point explicit in a Scholium appended to his paper, where he argued, in part:

That the same rule is the proper one to be used in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, seems to appear from the following consideration; viz. that concerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. For, on this account, I may justly reason concerning it as if its probability had been at first unfixed, and then determined in such a manner as to give me no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. But this is exactly the case of the event M . (1763, pp. 392-393)

Nearly all commentators have taken Bayes to mean that, confronted with an "unknown event," when we know nothing about its probability, we may assume all values between 0 and 1 equally likely—as if the probability of that event had been determined by tossing a ball on the unit table. Stigler (1982) has recently argued that Bayes was not in fact referring to a uniform distribution of probability for an unknown event. Rather, for him, an unknown event was one where all $n + 1$ possible outcomes were equally likely, from 0 occurrences of M to n occurrences. I think a close reading of the Scholium does indeed support this interpretation. Bayes then, according to Stigler, inferred a uniform distribution for the probability of p from the uniform distribution for the probability of the number r of events M . In fact, Stigler points out, a uniform distribution for r is a necessary, but not quite a sufficient, condition for a uniform distribution of p . The whole distinction is a rather small one, but Stigler credits Bayes with having at least spoken of a uniform distribution for the observed variable r rather than the unobservable variable p .

As in the case of Bernoulli, most commentators have assumed that it was doubt about the argument from analogy in his Scholium that made Bayes reluctant to publish his paper; and Stigler (1986) suggests, again, that the reasons for his hesitation were not philosophical but mathematical. For large values of n and r , where the theorem might be empirically useful, evaluation of the necessary expression was prohibitive. For one illustrative case, Bayes (and Price after him) tried to calculate upper and lower bounds for the probability of p lying within a certain interval, but their extraordinary effort still yielded only a uselessly rough estimate.

⁵ It was more careful than he is sometimes given credit for. Hogben (1957) takes him sharply to task for having real balls come to rest on Euclidean points, and he modifies the apparatus to have the balls falling into discrete holes in the table. But Bayes was scrupulous in speaking of areas rather than geometric points, as when he says, in a passage quoted by Hogben himself (p. 119n), "There shall be the same probability that it rests upon any one equal part of the plane as another" (Bayes, 1763, p. 385).

If Bayes was discouraged by the practical mathematical difficulties, Price's enthusiasm for the philosophical possibilities was hardly dampened. Causes may have become unknowable in the new methodology of the seventeenth century, but if they could no longer be proved, they could at least be made now the object of probability statements. All that was required for application of Bayes' rule was that we be utterly ignorant about the probability of the various possible causes *a priori*. This state of perfect ignorance is Foucault's example of genesis, my example of agenesis, from Chap. 2. The quotation is from Price's appendix to Bayes (1763).

Let us imagine to ourselves the case of a person just brought forth into this world and left to collect from his observation of the order and course of events what powers and causes take place in it. The Sun would, probably, be the first object that would engage his attention; but after losing it the first night he would be entirely ignorant whether he should ever see it again. He would therefore be in the condition of a person making a first experiment about an event entirely unknown to him. But let him see a second appearance or one *return* of the Sun, and an expectation would be raised in him of a second return, and he might know that there was an odds of 3 to 1 for *some* probability of this. This odds would increase, as before represented, with the number of returns to which he was witness. But no finite number of returns would be sufficient to produce absolute or physical certainty. (p. 409)

When Price speaks of “an odds of 3 to 1 for *some* probability” of a second return, he evidently means that the odds are 3 to 1 that the antecedent probability of the sun returning is greater than $\frac{1}{2}$ ⁶:

$$P\left(\frac{1}{2} < p \leq 1\right) = \frac{\int_{\frac{1}{2}}^1 p^1 (1-p)^0 dp}{\int_0^1 p^1 (1-p)^0 dp} = \frac{3}{4}.$$

(If we count appearances rather than returns of the sun, so that $r = 2$ instead of 1, the odds are 7 to 1.)

For all its obvious promise as a philosophical and scientific tool, Bayes' rule failed to attract much notice. The reasons for this surprising neglect are not altogether clear. Apart from the mathematical intractability for practical problems pointed out by Stigler, the only ready explanation lies in the impenetrability of Bayes' prose style and mathematical argument, a feature which few readers have failed to comment on. The consequences for subsequent development of statistical inference were rather slight, however, since Laplace soon produced a model which did capture the imagination of nineteenth-century philosophers of science.

⁶Dale (1982) argues that the correct solution is

$$\frac{\int_0^1 x^2 dx}{\int_0^1 x dx} = \frac{2}{3}.$$

This is Laplace's solution (see below) for the probability that the next day will have a sunrise, given that the first one did. But that is a different problem from the one Price solved.

4.2 Laplace

Laplace's philosophical intentions were explicitly signaled in the title of his “*Mémoire sur la probabilité des causes par les événements*” of 1774, written when he was only 25. Evidently inspired by the lottery instituted by the French government in the latter half of the eighteenth century, he posed the following problem:

If an urn contains an infinity of white and black tickets in an unknown ratio, and if $p + q$ tickets are drawn of which p are white and q are black, we ask the probability that in drawing a new ticket from this urn it will be white. (Laplace, 1774/ 1891, p. 30)

Laplace reasoned essentially as follows: We know nothing of the actual constitution of the urn (except that it contains at least p white and q black tickets), but we do know that different constitutions will give rise to our particular sample of $p + q$ with different probabilities. To each of the infinite number of possible values x of the probability of drawing a white ticket, there corresponds a different constitution of the urn. Laplace argued that in the absence of knowledge favoring one ratio of black to white over another, we may assume them all equally likely.⁷ This assumption seemed to be a simple instance of the same principle by which elementary probabilities in gambling were derived. In the absence of evidence for bias, we assume all six sides of a die or both faces of a coin equally likely to turn up. This principle was named by Laplace the principle of insufficient reason, to compare and contrast it with Leibniz' principle of sufficient reason.

It was meant by the latter principle that causes identical in all respects have always the same effects. On the other hand, if it is known only that the causes are alike in some respects, whereas their likeness or difference in other respects is unknown, the reason for expecting the same effect from all is insufficient. Alternatives become possible and probability replaces certainty. (Cox, 1961, pp. 102-103)

Keynes (1921/1973) rechristened the rule the principle of indifference.

In Laplace's urn problem, the principle amounted to replacing a *single* urn of *unknown* constitution with an *infinity* of urns of *known* constitutions, one of which is drawn at random. If we grant him that substitution, the solution to the problem he posed is straightforward: the probability that the next ticket will be white is

$$\frac{\int_0^1 x^{p+1} (1-x)^q dx}{\int_0^1 x^p (1-x)^q dx} = \frac{p+1}{p+q+2}.$$

⁷The difference between equiprobable *ratios* of black to white and equiprobable *constitutions* of the urn is obscured when the number of tickets is infinite but is obvious when there are only, say, two tickets in the urn. In this case there are four possible constitutions of the urn—BB, BW, WB, WW—each with probability $\frac{1}{4}$; but there are three possible ratios of black to white—1, $\frac{1}{2}$, 0—each with probability $\frac{1}{3}$. Keynes (1921/1973) points out that in Laplace's day it was conventional to assume all possible ratios equally likely, whereas in the twentieth century, we take all possible constitutions equally likely. The issue ultimately appears to be an empirical one (see Chap. 6).

This result was to be named by Venn (1888) the Rule of Succession. As may be apparent at once, its interest lay in application to the skeptical problem of induction. Adolphe Quetelet, whose contributions will be studied in the next chapter, is famous for his dictum, “*L’urne que nous interrogeons, c’est la nature*” (1846, p. 31). Karl Pearson, carrying the metaphor into the twentieth century, referred in *The Grammar of Science* (1892) to the “nature bag” (p. 176). The meaning of that metaphor was perhaps most clearly stated by Jevons (1874):

Nature is to us like an infinite ballot-box, the contents of which are being continually drawn, ball after ball, and exhibited to us. Science is but the careful observation of the succession in which balls of various character present themselves; we register the combinations, notice those which seem to be excluded from occurrence, and from the proportional frequency of those which appear we infer the probable character of future drawings. (p. 150)

(It is not clear why Jevons would have had voting done with balls rather than Laplace’s tickets. He claimed to be following Poisson’s example, but Poisson used an urn.)

Laplace’s urn, with its arbitrary proportions of black and white tickets, was a plausible model for inductive generalization with regard to characteristics as superficial as the color of lottery tickets. We might almost as easily imagine drawing a sample of black or white swans from the bag as black or white tickets. If I have seen 100 swans in my life, and all were white, the probability that the next one I saw would be white should be 101/102. Jevons (1874), picking up an example from de Morgan (1838), thought that if he had watched 12 steamboats going down the river, all with a flag, he could lay odds of 13 to 1 that the next would be flying a flag also. It seems not to have occurred to him that, if all the flags had been red, he could have laid the same odds that the next boat would have a red flag—which seems to rule out the possibility of a boat having a green flag.

The most flamboyant, and celebrated, example was Laplace’s own. It was given, not in his 1774 memoir, but in his popular *Philosophical Essay on Probabilities* in 1795. By this time he had seen Bayes’ essay, with Price’s implicit reference to Hume (or Butler) in the sunrise example, and offered his own response. Imagine, said Laplace, the world to have been created 5000 years ago.⁸ Allowing for leap years, that makes 1,826,213 days, all of which have been blessed with a sunrise. (Set aside the question of what a day would be without a sunrise.) Hence the probability that the sun will rise tomorrow is 1,826,214/1,826,215.⁹

⁸Laplace was actually referring to recorded history, since if there had been days without sunrises before then, we wouldn’t know about them. His figure is, in any event, not so different from that accepted by some of his contemporaries, like Bossuet, for the actual creation of the world: In 1650, Archbishop James Usher calculated that the world was created in 4004 B.C., on October 26, at 9 o’clock in the morning (Kearns, 1979). (Whether the archbishop intended Greenwich or Eden time I am not sure.) Quetelet (1846), repeating Laplace’s example, used Usher’s figure.

⁹Laplace went on to note Buffon’s solution to the problem, which dates from around the same time. Buffon held that, since a day either has a sunrise or it does not, the antecedent probability of its having a sunrise is $\frac{1}{2}$, and that this figure is not to change as a result of experience. Consequently, for him, the probability of a sunrise after a string of n sunrises has been observed is $1 - \frac{1}{2^n}$.

4.3 Criticism

Whatever plausibility the Rule of Succession could claim as a tool of inductive inference derived from application to relatively “detachable” characters like the presence of a flag on a steamboat or the color of a swan. Yet attributes could also be too obviously superficial for the Rule to have any plausibility. Keynes (1921/1973) recalls that Peirce took the first five poets from a biographical dictionary and found several mathematical properties common to their ages at death (e.g., the difference of the two digits, divided by 3, leaves a remainder of 1). We would not rationally give any credence to such a law even if it were confirmed for 100 cases, astonishing though such a result would be.

There was nothing in the formulation which restricted its application to surface characters, and for generalizations of more scientific interest, the application was hardly less strained. The Rule of Succession would hold, with equal authority, that, given the same sample of swans, the probability is 101/102 that the next swan should have feathers, or a head. If I had never seen a swan in my life, the probability that the first I saw should be white or feathered, etc., would be $\frac{1}{2}$, and no more.

Indeed, with the concept of causality banished to metaphysics, there was no scientific way to distinguish surface from deep, or extrinsic from intrinsic, characters. With Aristotelian causes no longer to bind things together, entities became arbitrary aggregates of characteristics, to be studied by plotting joint distributions, distances between genera in multidimensional space, and so on. The mathematics for such dreams was not yet worked out, but that was to prove indeed a merely technical problem.

It was the loss of organic connection that created the skeptical problem of induction and gave it such urgency and preeminence in the philosophy of science from the mid-eighteenth century on. And it was for such a dissociated world that the Rule of Succession was made. If certainty could not be attained, with nothing but historical frequency or habitual association to support the belief that the sun will rise tomorrow or that the next piece of bread I eat will nourish me, what the Rule of Succession offered in its stead was a very precise quantification of uncertainty.

I have described this mathematical formalization of the world as a more or less necessary consequence of the disappearance of an organic world view, as all that was thereafter possible; but it is also true that the mathematical model, once established, was in turn antagonistic to organicism. The conditions for successful application of the mathematical formulas are invalidated, or at least complicated, by any special knowledge of individuals or relationships proceeding from the latter. The Rule of Succession, in particular, required for its application a perfect ignorance of possible causes of a phenomenon.

It can thus be seen how aptly the new concept of statistical inference mirrored the Continental epistemology of Condillac and the Idéologues, as well as the British associationist psychology of Hartley and Locke. All assumed, or required, a tabula rasa start and were concerned to show how knowledge could then arise merely

through the combination of elements and the repetition of patterns. It is perhaps also clear how the procedures of Bayes and Laplace qualify for the title of statistical inference. Whether we are concerned to infer backward from observed events to unknown causes or laterally from one sample of members of a class to the next, we are involved with a problem of inference. But the probability statements we make on the basis of these procedures are not merely epistemic or vaguely quantified probabilities; they are aleatory, derived from chance mechanisms. French philosophy was eventually to become much more sophisticated, but the atomic program of the Idéologues would be taken over by logical positivists and their relatives at the beginning of the twentieth century, so that the associationist psychology and the epistemology, and their formalization in statistical inference, would all remain strong 200 years later.

The classical theory of statistical inference thus owed its success to its supporting role in this larger philosophical and psychological framework, as well as to its great simplicity and power. In inductive generalization, nothing was beyond the reach of the Rule of Succession. Any unknown probability whatever could be handled by assuming all values between 0 and 1 equally likely. Wherever an unknown probability occurred in a formula, it could be replaced with a definite integral, and an exact numerical result derived. Total ignorance was effortlessly converted into definite knowledge. Indeed, it might well be said of the Rule of Succession, as was once said of Leibniz' pre-established harmony, that it was “*une idée trop ingénieuse, trop recherchée, pour être vraie*” (“Fatalité,” 1756, p. 423).

More specifically, the Rule recalls Lalande's delicate commentary on Lachelier's *Du Fondement de l'Induction*: “*On a plus souvent l'occasion de l'admirer que de l'utiliser*” (quoted by Piaget, 1950/ 1974, p. 195). For, in fairness to Laplace, it must be said that it is not clear how seriously he took the Rule of Succession himself. It did not appear in his major work on the subject, the *Théorie Analytique des Probabilités*, first published in 1812, and he did not use it in his own empirical work in astronomy. The sunrise example occurs only as a casual comment in the non-mathematical *Essay*, and even there, Laplace followed it up with a more traditional appeal to causation: “But this number [i.e., odds of 1,826,214 to 1] is incomparably greater for him who, knowing through the totality of phenomena the regulative principle of the days and seasons, sees that nothing in the present moment can arrest its course” (1816, p. 23).

If the extraordinary success—in philosophy, not in science—of the Rule of Succession was due in large measure to its fit with the prevailing epistemology, it is still remarkable that an idea with so little intrinsic merit should have so thoroughly captured the imagination of a century and a half. Some specific features of the concept of probability may help further to explain that success.

In the first place, the fact that probability was understood as relative to knowledge tended to blur the distinction between known and unknown probabilities. The double uncertainty of the latter was not clearly seen as requiring a separate marker; uncertainty was uncertainty. Thoroughgoing subjectivists in the twentieth century would deny meaning to the concept of an unknown probability altogether: It implies that we don't know our own state of belief. De Finetti (1974) makes an analogy with

“thinking that in a statistical survey it makes sense to indicate, in addition to those whose sex is unknown, those for whom one does not even know ‘whether the sex is unknown or not’” (p. 84). In modern terms, known probabilities are associated with direct problems, as part of the hypothesis; unknown probabilities are associated with inverse problems, as the quantity sought. Hence the distinction between these two types of problems could easily be lost; straightforward, noncontroversial assumptions were confused with outrageous ones.

A second factor was no doubt the difficulty in distinguishing directional implications. Direct problems concern the probability of outcomes, given some hypothesis; inverse problems ask the probability of the hypothesis, given the outcome. Many present-day psychologists have trouble with such distinctions in the abstract, as we saw in Chap. 1. The connection between a priori probability and long-term relative frequency is close enough to lend itself to a careless assumption of bidirectionality—again collapsing the distinction between ordinary probability problems and statistical inference.

This second difficulty was exacerbated by the lack of standard notation for conditional probability until the twentieth century. In his history of inverse probability, Dale (1991) finds a number of instances where authors apparently forgot the conditional character of the probabilities they were calculating and proceeded to unwarranted claims. Hacking (1980) remarks of the history of probability theory more generally, “No one was able properly to conceptualize what had been discovered. Even when the formalism could be written down with some rigor, no one knew what it meant” (p. 115).

The success of the Rule of Succession is attested by the fact that twentieth-century authors still felt called upon to mount serious criticism of it. C. D. Broad (1918, 1920), and Jeffreys (1961) following him, argued, for example, that the Rule of Succession is biased against homogeneity of a population. Suppose there are N elements of a class to which we wish to generalize; the question concerns the probability that a given member (e.g., the next to be observed) will possess a certain characteristic, and that probability is just the relative frequency of occurrence of the characteristic in the class. In the Laplacean approach, in the absence of knowledge concerning the value of that proportion, we may assume all possible values equally likely. There will be $N + 1$ such possible values, each thus assigned a prior probability of $1/(N + 1)$. Broad pointed out that if N is large at all, which is ordinarily the case, then it would take an enormous number of observations to support any particular value over the others. In particular, Jeffreys (1961) argues, the formula would make the (frequently occurring) extreme values, 0 and 1, so incredible that no realistic amount of data would support them. This criticism is the point of departure for Jeffreys' own version of Bayesian theory (Chap. 8).

Some of the most telling arguments against the Rule of Succession were made by Keynes (1921/ 1973), who argued that the Rule is self-contradictory. The binomial theorem on which it is based pertains to what are now known as Bernoulli trials, i.e., independent trials with a fixed probability of success on each. But on the Rule's own showing, that probability changes from trial to trial. Thus after an event was observed once, its probability of happening again, according to the Rule, is not

$1/2 \cdot 1/2$ but $1/2 \cdot 2/3 = 1/3$; the probability of its happening N times in succession is not $1/2^N$ but $1/2 \cdot 2/3 \cdot 3/4 \cdots \cdot 1/(N+1) = 1/(N+1)$. The Rule of Succession assumes that the observations are independent and that the probability is constant from one to the next, then proves that the trials are dependent and that the probability constantly changes.

Perhaps the finest rejoinder to Laplace was that made in 1891 by Karl Bobek, who showed, using the same assumptions as Laplace, that the probability that the sun will rise every day for the next 4000 years is only about $2/3$ —“a result less dear to our natural prejudices” (Keynes, 1921/1973, p. 418).

From the point of view of ultimate impact, overwhelmingly the most important criticism of the Rule of Succession focused on the principle of indifference, which Laplace had applied to unknown probabilities. This objection derived from an empiricist critique of the definition of probability, and provided one impetus for development of the frequency theory of probability, which is the subject of Chap. 6. A new theory of statistical inference based on the frequency theory of probability would also be based on the normal rather than the binomial distribution, and development of normal distribution theory in the social sciences is the subject of Chap. 5.

The importance of the criticism of the principle of indifference was thus that it ultimately led to a new theory of statistical inference. The other criticisms, however logically devastating, were ineffectual. There was a growing uneasy awareness, to be sure, that there was something peculiar about a scientific principle which required for its legitimate application not any specific information but a very precise balance in our ignorance of alternatives; that the Rule of Succession proved altogether too much; that it bought its impressively exact numerical results with assumptions as groundless as they were extravagant; that change by imperceptible increments does not adequately represent the growth of our knowledge; and so on. But the sensible restraint and common sense of the few who were urging these criticisms left them by and large out of the analytic ambitions of the nineteenth and twentieth centuries. The decline in popularity of the Law of Succession among probability theorists had more to do, in other words, with the fact that they were now busy fashioning new tools for tackling the same kinds of problems, than with recognition of the limitations of the old methods.

References

- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.
- Bellhouse, D. R. (2004). The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. *Statistical Science*, 19, 3–32.
- Broad, C. D. (1918). On the relation between induction and probability. Part I. *Mind*, 27, 389–404.
- Broad, C. D. (1920). On the relation between induction and probability. Part II. *Mind*, 29, 11–45.
- Cox, R. T. (1961). *The algebra of probable inference*. Baltimore: Johns Hopkins University Press.
- Dale, A. I. (1982). Bayes or Laplace? An examination of the origin and early applications of Bayes’ Theorem. *Archive for History of Exact Sciences*, 27, 23–47.

- Dale, A. I. (1991). *A history of inverse probability from Thomas Bayes to Karl Pearson*. New York: Springer-Verlag.
- de Finetti, B. (1974). *Theory of probability* (Vol. 1). New York: Wiley.
- de Morgan, A. (1838). *An essay on probabilities, and on their application to life contingencies and insurance offices*. London, UK: Longman, Orme, Brown, Green, and Longmans, and John Taylor.
- Fagot, A. M. (1980). Probabilities and causes: On life tables, causes of death, and etiological diagnoses. In J. Hintikka, C. D. Gruender, & E. Agazzi (Eds.), *Proceedings of the 1978 Pisa Conference on the History and Philosophy of Science* (Vol. 2, pp. 41–104). Dordrecht: Reidel.
- Fatalité* [Fatality] (1756). In D. Diderot & J. le R. d'Alembert (Eds.), *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers* [Encyclopedia, or reasoned dictionary of sciences, arts, and crafts] (Vol. 6, pp. 422–429). Paris, France: Briasson, David, le Breton, et Durand.
- Hacking, I. (1980). Grounding probabilities from below. *Philosophy of Science Association*, Vol. I, eds. R. Giere and P. Asquith.
- Hacking, I. (1990). *The taming of chance*. Cambridge, UK: Cambridge University Press.
- Hogben, L. (1957). *Statistical theory: The relationship of probability, credibility and error*. New York: Norton.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon Press. (1st ed., 1939).
- Jevons, W. S. (1874). *The principles of science*. London, UK: Macmillan.
- Kearns, E. J. (1979). *Ideas in 17th-century France: The most important thinkers and the climate of ideas in which they worked*. Manchester: University of Manchester Press.
- Keynes, J. M. (1973). *A treatise on probability*. New York: St. Martin's Press. (Original work published 1921).
- Laplace, M. Le Comte. (1816). *Essai philosophique sur les probabilités* [Philosophical essay on probability] (3rd ed.). Paris, France: Courcier.
- Laplace, P. S. Marquis de (1891). Mémoire sur la probabilité des causes par les événements [Memoir on the probability of causes from events]. *Mémoires de l'Académie Royale des Sciences de Paris*, 1774, 6, 621–656. (Reprinted in *Oeuvres complètes*, Vol. 8, pp. 27–65).
- Pearson, K. (1892). *The grammar of science*. London, UK: Walter Scott.
- Pearson, K. (1978). The history of statistics in the 17th and 18th centuries, against the changing background of intellectual, scientific, and religious thought. E. S. Pearson Ed.). London, UK: Griffin.
- Piaget, J. (1974). *Introduction à l'épistémologie génétique*. Vol. 2: *La pensée physique* [Introduction to genetic epistemology. Vol. 2: Thinking about physics] (2nd ed.). Paris, France: Presses Universitaires de France. (1st ed., 1950).
- Quetelet, A. (1846). *Lettres sur la théorie des probabilités* [Letters on the theory of probabilities]. Brussels: Hayez.
- Shafer, G. (1982). Bayes's two arguments for the rule of conditioning. *Annals of Statistics*, 10, 1075–1089.
- Stigler, S. M. (1982). Thomas Bayes' Bayesian inference. *Journal of the Royal Statistical Society (Series A)*, 145, 250–258.
- Stigler, S. M. (1983). Who discovered Bayes' Theorem? *American Statistician*, 37, 290–296.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
- Venn, J. (1888). *The logic of chance* (3rd ed.). London, UK: Macmillan. (1st ed., 1866).

Chapter 5

Nineteenth-Century Developments in Statistics



5.1 Descriptive Statistics

The Bayes-Laplace theory of statistical inference survived, criticized but unchanged, into the twentieth century. The new theory that appeared then would build upon two sets of intervening developments: concepts and methods of descriptive statistics growing out of error theory in astronomy, and further reflection on the concept of probability (to be taken up in Chaps. 6 and 8). The developments in descriptive statistics that are the subject of this chapter essentially span the nineteenth century, from Laplace to Karl Pearson, though one root extends back farther: the derivation of the normal curve from the binomial.

The conceptual arrangements that resulted from these developments were, like the duality of the concept of probability, to become so thoroughly absorbed into the common sense of the twentieth century that it is a challenge for us to imagine a world which lacked these particular constructions. Yet what are to retrospect even the slightest advances were, in their own time, breakthroughs or radical departures from existing thought, which took a succession of thinkers over many decades to consolidate. The first such advance was the application of probability theory to errors of measurement in astronomy; the second was the application of the normal law of error to natural populations. A third, which was building to some extent parallel to these two, was what Porter (1986) refers to as “statistical thinking”: recognition of a comprehensibility in aggregates which is denied to individuals.

5.1.1 The Normal Curve in Astronomy

The discovery of the normal curve was to have a curious history in itself. For over a century that discovery was credited to Laplace and Gauss, who built on the work of Lagrange, Legendre, Euler, and Stirling; indeed, the curve is still sometimes referred to as a Gaussian distribution. In 1924, Karl Pearson, examining a copy of Abraham de Moivre's *Miscellanea Analytica* of 1730, noticed a seven-page supplement bound into the back of the book; this supplement, in Latin, presented the derivation of the normal curve from the binomial. It was written in 1733 and was bound into the few remaining unsold copies of the *Miscellanea Analytica*. It hence remained extremely rare—only seven copies have ever been found, and one of these, in Berlin, was lost during World War II—and was not known to de Moivre's successors; but, following Pearson's discovery, de Moivre was belatedly awarded the credit. The real irony, however, as David (1962) points out, is that the now-famous supplement was also included in the second edition of de Moivre's *The Doctrine of Chances* in 1738 and the third edition in 1756. This book was widely read, yet no one ever noticed that de Moivre had preceded Laplace and Gauss by 50 years. De Moivre was clearly a mathematician of considerable talent; there have always been those, starting with de Moivre himself, who felt that he was insufficiently appreciated by his contemporaries.

When the normal curve reappeared for good, as it were, nearly a century after de Moivre, it was in the context of a sophisticated and general theory of error in astronomical measurements. It was perhaps natural that the phenomenon of measurement error should first have been of concern in astronomy, the most exact of sciences; the appearance of this concern also coincided with a period of rapid advances in the technology of measuring instruments, which focused interest, almost paradoxically, on their limitations. At the limit of precision of any measuring process, the observations will vary randomly. Though we take it for granted today that the way to deal with this random variation is to strike an average, the obviousness of this procedure can only be said, historically, to be an acquired self-evidence. Variation in measurements of the same quantity entails the presence of error, and, to many astronomers of the eighteenth century, it was hard to see the advantage of averaging in values presumed to be erroneous. The discovery of the “personal equation” is credited to the Astronomer Royal Nevil Maskelyne in 1796; but when Maskelyne noticed that the observations made by his assistant David Kinnebrook were systematically diverging from his own, he didn't average the two sets; he fired Kinnebrook (Stigler, 1986). In general, the focus of mathematicians, then as now, was more naturally on the maximum possible error than on the most likely error. As observations subject to error are aggregated, the error bounds do indeed increase, even as the expected error decreases; hence there was little inclination to think in terms of averaging observations.

The development of the normal least-squares theory of measurement error, as characterized by Stigler (1986), was long and complex, with many conceptual shifts and technical advances necessary; there were also many paths that were abandoned.

Closely related to the problem of whether and how to increase accuracy by combining observations was the question of how multiple observations of some phenomenon might be distributed. The latter question brought in the nascent probability theory. A contribution by Simpson, small in itself but significant in its implications, points up the link between these two issues. Simpson was simply the first to frame the problem in terms of the distribution of errors rather than the distribution of the observations themselves; and this shift, Stigler suggests, facilitated thinking in terms of averages. Simpson's own illustrative error curve was triangular, sloping away symmetrically from a most likely value to fixed upper and lower limits. In a symmetrical distribution, obviously, the errors average out to zero.

The successive contributions of Laplace are especially interesting. He was a major contributor to both probability theory and astronomy, and he had derived the normal curve in at least two different contexts. One was the distribution of the ratio of black to white tickets in the urn, after the sample had been drawn (see Chap. 4); the other was in the proof of the Central Limit Theorem: that curve is the limiting distribution of sums (or means) of independent and identically distributed random variables. But it did not at first occur to Laplace to consider the normal distribution as a possible error curve. Approaching the problem from the standpoint of probability theory, he looked for distributions that could be derived on the basis of some application of the principle of insufficient reason. These efforts ended in mathematically intractable curves; they are admirably traced by Stigler (1986).

It was Gauss, in his *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium* of 1809, who argued for a normal distribution of errors. Stigler (1986) characterizes his argument as boldly circular: Gauss proved that if the mean of the observations yields the most probable value, then the errors must be normally distributed; hence, since the mean was by then generally acknowledged as the way observations should be combined, the error curve should be assumed normal. This bit of reasoning might not have had any greater impact than Simpson's argument for a triangular distribution or Daniel Bernoulli's (1977/1961) for a semiellipse, were it not for Laplace. Having just proved the Central Limit Theorem, Laplace was immediately struck by the possibility of regarding errors as aggregates in themselves: That structure would provide a noncircular rationale for the normal distribution. It also fit with the normal curve as an approximation to the binomial.

The normal model assumed one ideal, true value, from which any given measurement represented a certain deviation. It was further assumed that each deviation resulted from a multitude of very small “elementary errors”—distortions produced by momentary changes in the atmosphere, irregularities in a cogwheel of the telescope, and so on. Each individual factor was conceived as a Bernoulli “trial,” with a certain fixed probability of a positive or negative effect on the resultant observation; the factors, or trials, were assumed to be independent of each other and to combine randomly to yield a given deviation. Provided that the assumptions of independence and random combination hold, and further that the number of such elementary factors is very large, that their effects are approximately of equal magnitude, and that negative and positive effects occur with equal frequency, the distribution of resultant deviations will tend to take on a normal form. (With subsequent sophisticated

developments in the mathematical theory, certain of these conditions—e.g., the assumption of equal magnitude of elementary errors—could be relaxed somewhat.)

The normal curve did not go unchallenged as a law of error. Unlike the distributions previously proposed, it placed no bounds whatever on the magnitude of possible errors. Daniel Bernoulli (1777/1961) criticized this feature as flatly unrealistic. At the same time, the normal curve tended to be too conservative with respect to large errors. Comparisons with empirical data by Bessel in 1818 and others indicated that the occurrence of such errors tended to be underestimated by the normal law. But the combination of a theoretical rationale and surpassing mathematical convenience insured that the search for models of error ended with the normal distribution.

5.1.2 Application of the Normal Distribution to Populations: Quetelet

The next step in the evolution of normal theory involved a transplanting across disciplines and thus required an individual familiar with both. Adolphe Quetelet was well situated for the task. A Belgian who was trained as an astronomer and studied the methods of Laplace, he was also, along with Auguste Comte, one of the founders of modern sociology. In fact, his talents were wider yet, as Keynes (1921/1973) somewhat scornfully acknowledges:

Before accepting in 1815 at the age of nineteen (with a view to a livelihood) a professorship of mathematics, Quetelet had studied as an art student and written poetry; a year later an opera, of which he was part-author, was produced at Ghent. The character of his scientific work is in keeping with these beginnings. There is scarcely any permanent, accurate contribution to knowledge which can be associated with his name. But suggestions, projects, far-reaching ideas he could both conceive and express, and he has a very fair claim, I think, to be regarded as the parent of modern statistical method. (p. 334)

Whether Keynes himself, with these disparaging remarks, was responsible or not, Hogben (1957) is right in noticing that it was just at that time that acknowledgement of Quetelet's role abruptly disappeared. A century ago his name was among the most illustrious in the field; it is hardly known today.

The normal model in astronomy applied to a large number of observations of the same putative quantity: an ideal, true value from which all actual measurements represented deviations or errors. Quetelet's distinctive, and enormously consequential, contribution was to apply the normal curve to measurements of different individuals. He drew the analogy in explicit terms: If we construct a distribution of heights of men, for example, he said:

The numbers will present themselves exactly as if they were the result of measurements taken on one and the same person, but with instruments crude enough to justify the size of the variations. If there does not exist a constant cause of error, it will happen in fact that after a large number of measurements, the positive deviations will balance the negative, in such a way that the mean will give the true height that one sought to determine; one will

even find that the divers results obtained, when arranged in order of magnitude, will fall symmetrically on the two sides of the mean. (Quetelet, 1848, p. 18)

Quetelet was a child of his time; it was just in the 1830s that governments, in an effort to “see like a state,” in Scott’s (1998) phrase, began to compile, often from conscripts, the sort of datasets that would come to constitute subject matter of the discipline. Quetelet obtained, for example, the chest measurements of 5738 Scottish soldiers (presumably made by a tailor; Stigler, 1986) and fitted a binomial distribution; he found the probable error (the term introduced by Bessel for the median deviation) to be 33 millimeters. “I now ask,” said Quetelet (1846),

if it would be an exaggeration to claim it an even bet that someone engaged in taking measurements of the human body would err by about 33 mm in measuring a chest of more than a meter in circumference? Well, accepting this probable error, 5738 measurements taken on the same person would certainly not group themselves with more regularity (as regards to order of magnitude) than the 5738 measurements taken of the Scottish soldiers. And if we were given the two series of measurements without being told which was which, we would be at a loss to tell which series was taken on 5738 different soldiers, and which had been obtained on one and the same person, with less ability and rougher judgment. (p. 137)

With some of his other empirical distributions, the analogy became more strained. In fitting the distribution of 100,000 French conscripts, he took the report by one Birch of an individual 17 inches tall at the age of 37, or 3 feet 11 inches below the mean of 5 feet 4 inches. He then derived his upper limit for human height by taking the same deviation in the other direction. The resulting limit of 9 feet 3 inches was never attained, to the best of Quetelet’s knowledge; the greatest authenticated height he found was 8 feet 3 inches, of a Swedish bodyguard of Frederick the Great.¹ As Hankins (1908) was later to remark,

Quetelet’s analogy between the efforts of nature in producing a type and of man in measuring a height of a person was in appearance considerably weakened by the fact that an error of three feet eleven inches in measuring a height of five feet four inches is extremely improbable. (pp. 519–520)

Quetelet (1846), however, drove his point home further with an analogy with the many reproductions that had been made of an ancient statue. These copies, he noted, were not all exact, and the variation among them paralleled the variation among live persons.

The most remarkable and interesting aspect of Quetelet’s thought here is that he did *not* see himself as making a radical conceptual shift, but merely as finding a broad new domain of application for what was later called the normal law of error. He regarded his achievement, on the contrary, as vindication of his position that all the sciences were united, or should be, by the same methods. The issue was a live one because Comte, from whom Quetelet borrowed the term *physique sociale*

¹One unexpected finding to emerge from his labors was a subtle departure from normality in the vicinity of 1.57 meters: a surplus of observations just below that value, a corresponding deficit just above. The evident implication was that 2275 young Frenchmen had bent their knees to avoid the draft (Tankard, 1984).

(leading Comte to abandon it), argued at the same time for a hierarchy of sciences, with methodological distinctions between the levels. Quetelet was satisfied to have shown that the binomial error law of astronomy was applicable to the social sciences.

As applied to the domain of natural variation, the “binomial law,” as Quetelet called it, implies that any given measurement results from a myriad of very small “accidental” errors. Natural variation is conceived as the product of chance processes, of Bernoulli trials. If we think, following Laplace, of an urn containing millions of tiny tickets, each inscribed with an error of a certain size and direction, then an individual’s height (or moral stature, etc.) is determined as if a large scoopful of these tickets had been drawn from the urn and the errors indicated thereon algebraically summed. Whatever the dimension of variation, the norm was to be discerned by purifying actual observations of accidental error through the resort to large numbers.

We must lose sight of man taken in isolation, and consider him only as a fraction of the species. In stripping him of his individuality, we shall eliminate all that is merely accidental; and the individual peculiarities which have little or no influence on the whole mass will be effaced of their own accord, and allow us to grasp the general results. (Quetelet, 1835, v. 1, pp. 4-5)

A major consequence of the metaphor of errors of measurement in astronomy was to focus attention sharply on the mean of the distribution as the underlying true value. For Quetelet, natural variation represented Nature’s errors, as it were, in her striving to produce the norm—essentially a failure of quality control at the very highest level. This norm Quetelet saw as the ideal type, which he called *l’homme moyen*. With this concept he really put the ancient doctrine of perfection in moderation on a modern statistical footing. Being free from error, and typical in all proportions, *l’homme moyen* was, as Quetelet conceived him, the model of physical beauty. Quetelet seemed even to imagine that *l’homme moyen* would represent an ideal in all respects that might be measured. Stretched as the notion was in the case of physical beauty, it is even more difficult to imagine mediocrity as ideal in the case of moral characteristics.

An ardent collector of social statistics, Quetelet was profoundly impressed by the statistical regularities which kept turning up wherever he looked; and he made endless classifications of his data (all, curiously to a modern eye, univariate, Stigler notes), to reveal *l’homme moyen* in all his dimensions of variation.² He announced, for example, on finding that more crimes were committed by men in their 20s, that the *average man* has the greatest propensity for crime while in his 20s—whatever

²Classification schemes of course vary across cultures. One of the more interesting examples is a frequency distribution of occupations which Quetelet (1835/1969) quotes from the *Rapport au Roi* of 1829. It starts with “individuals who work on the land, in vineyards, forests, mines, &c., 2453” and proceeds on down (*not* in order of decreasing frequency) to the seventh class, comprising “innkeepers, lemonade-sellers, servants, &c., 830; the eighth, artists, students, clerks, bailiffs, notaries, advocates, priests, physicians, soldiers, annuitants, &c., 449; the ninth, beggars, smugglers, strumpets, &c., 373” (p. 85).

“the inconsistency of viewing him as a type of perfection and nevertheless as somehow endowed with a propensity to crime” (Hankins, 1908, p. 517).

Striking examples in the area of moral statistics—e.g., the “frightening exactitude with which crimes reproduced themselves” (Quetelet, 1835, p. 10) across generations—inspired widespread debate about such issues as free will: If there were some law operating to maintain a constant rate of murder or suicide, then it seemed that some individuals were metaphysically constrained to commit these crimes just to keep the numbers up (cf. Porter, 1986). Quetelet’s (1848) notion of government was in fact that the state should intervene in human affairs just to restore equilibrium and statistical regularity when it was needed.

Along with the concept of the average man, Quetelet’s work illustrates a parallel nineteenth-century development in which he also played a major role: what Porter (1986) calls “the rise of statistical thinking.” The groundwork for the notion of the comprehensibility of aggregates had been Bernoulli’s proof of the lawfulness of large numbers. In the political realm, Hirschman (1977) goes so far as to suggest a certain cognitive motivation for the development of capitalism: If the whimsicalness and irrationality of people foiled understanding or prediction on the individual level, their “stubborn brutishness,” translated into the profit motive, could be used to achieve an intelligibility and predictability in the aggregate, through Adam Smith’s invisible hand principle. The predictability of the aggregate also figured prominently in the argument from design. Pearson (1978) credits William Durham with having originated the idea of God in stable statistical ratios; it was similarly the theme of Johann Peter Süssmilch’s celebrated *Die Götliche Ordnung* of 1741.³ But where Durham and Süssmilch saw statistical regularities as divine, Quetelet saw them as natural (Daston, 1987), and it was Quetelet more than anyone else who pursued the implications of the idea of statistical stability and established it in popular thought. The *particular* comprehensibility he found, of course, with his reification (not to say deification) of the mean, was his own and was criticized in his day even as it has persisted in attenuated form to the present. We may smile to read that the average American family has 2.3 children, but only because of the discreteness of that attribute; with respect to continuous dimensions, the “average man,” aka the man in the street, is very much a fixture of our thought.

Cournot (1843) was one of the first to criticize Quetelet’s conception of the average man, questioning whether a composite of average characteristics—height, weight, forearm length, and so on—would yield an average man, as Quetelet implied, or even a biologically viable being. By analogy, he pointed out that, if we averaged the lengths of the sides of many right triangles, the result would almost certainly not be a right triangle. A similar complaint was lodged by Bertrand (1889): Since some biological characteristics are nonlinearly related, not all of them can be normally distributed. Weight, being proportional to volume, varies more nearly with

³ As an argument from design, of course, it was not nearly so persuasive as the rhetorical question posed by Bernard Nieuwentyt about why God created the ocean: “If there were no ocean, how would ships get here from the Indies?” (quoted in Pearson, 1978, p. 300). It actually seems not to have occurred to Nieuwentyt that, if there were no ocean, we wouldn’t be taking a boat.

the cube of height than with height per se; thus if one variable is normally distributed, the other cannot be.

The physiologist Claude Bernard, to mention a final example, took issue with the assumption of averages per se as the *quaesita* of science. He insisted that “Averages are applicable only to reducing very slightly varying numerical data about clearly defined and *absolutely simple cases*” (1865/ 1952, p. 190). Otherwise, he argued, the process of averaging serves to obscure real and meaningful variation, which can guide us in the prosecution of causal inquiry.

If we collect a man’s urine over a period of twenty-four hours and mix all the samples to obtain an analysis of the average urine, we will have precisely an analysis of a nonexistent urine; because the urine during digestion differs from that during fasting, and the differences disappear in the mixture. The sublime example of this type was dreamed up by a physiologist who, having taken urine from the urinal of a railway station where people of all nations passed [*sic*], thought he could thus present an analysis of the *average* European urine! (p. 189)⁴

Bernard pressed his point with further examples, including the following:

A great surgeon performed several operations by the same procedure; he then made a statistical statement of cases of death and cases of cure, and he concluded, according to statistical method, that the law of mortality in this operation is two out of five. Well, I say that this relation means absolutely nothing scientifically and gives us no certainty in performing a new operation because we don’t know if this new case should be among the cures or the deaths. What should really be done, instead of empirically compiling facts, is to study them more exactly, each in its own special determinism. . . . There is evidently something that was the cause of death in the patient who succumbed, and that was not found in the patient who recovered; it is something which we must determine, and then we can take action on these phenomena or recognize and predict them exactly; only then will we have attained scientific determinism. But it is not with the aid of statistics that we shall reach it; statistics never has and never could tell us anything about the nature of phenomena. (pp. 192–193)

Bernard’s was a minority voice. For Quetelet’s contribution was not merely to make natural populations an object of mathematical investigation, nor merely to focus the attention of social scientists forever afterward on averages—both very considerable legacies; in a broader sense, he was the one most responsible for the success of a mathematical conception of the social sciences. Speaking of the calculus of probability, he wrote:

Everything which can be expressed numerically becomes its province; the more the sciences are perfected, the more they tend to enter its domain, which is a kind of center toward which they come to converge. One could even . . . judge the degree of perfection which a science has attained by the greater or lesser facility with which it is amenable to calculation, in accordance with that old saying which is daily confirmed: *mundum numeri regunt*. (Quetelet, 1828, p. 233)

Quetelet did not, of course, originate the idea of scientific perfection as mathematical, but he did give it vast new hopes. Like a modern progressive politician, it

⁴We can imagine what Bernard would have thought about a report (www.wtopnews.com/index.php?nid=25&sid=737418) that Fairfax County, Virginia, was testing wastewater for urinary byproducts of cocaine.

promised so much that the temptation to believe and go along was all but irresistible.

5.1.3 Galton, Pearson, and the Biometricians

For the future development of psychology, the decisive influence of Quetelet was on Francis Galton. According to Karl Pearson, his devoted biographer, Galton in later life tried to minimize Quetelet's influence; Galton himself, in his own autobiography, explains that he encountered Quetelet through a geographer, William Spottiswoode, who had crudely tried to fit the normal distribution to the direction of orientation of 11 mountain ranges (Stigler, 1986). Galton read the English translation of Quetelet's *Lettres sur la Théorie des Probabilités* and was quite as enchanted with the normal law of error as its Belgian champion in sociology.

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the “Law of Frequency of Error.” The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. (Galton, 1889, p. 66)

Galton shared Quetelet's fascination with numbers, with averages, and with the possibility of a quantitative science of man. When on two occasions he sat for a portrait, he estimated the number of brushstrokes in each to be about 20,000 (Galton, 1905); and, sitting near the back of the hall in a boring lecture, he counted the number of fidgets in the audience (Galton, 1885). He was the first to make composite photographs,⁵ thus bestowing on *l'homme moyen* a particularly vivid reality. It was evidently Galton's hope that there would be something useful in revealing the face of the average man, or the average criminal. Still, he deplored the narrow focus on means to which Quetelet's methods were already leading:

It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. (1889, p. 62)

He went on:

Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of man. (pp. 62–63)

⁵According to Lottin (1912/1969), the idea was suggested by Cheysson in 1886.

Galton found Quetelet's concepts useful in his studies of heredity, in which he is best known for discovering the phenomenon of regression. In one experiment with sweet peas, he found that seeds of any given size yielded offspring exhibiting a miniature bell-shaped distribution.⁶ Initially puzzled as to why variability then didn't continue to increase across generations, he noticed that the mean of each such subdistribution lay closer to the overall mean than did the parents of that particular set. This was the phenomenon he labeled reversion, and later regression to the mean.

In a subsequent study he assembled data on heights of parents and children into a bivariate table. Since women in his sample averaged 12% shorter than men, he multiplied mothers' heights by 1.08 to bring them up to standard, then struck an average between the two parents, which he called the "Mid-Parent." Porter (1986) suggests that it was Galton's previous experience with maps and isothermal lines which facilitated his recognition of elliptical contours of equal frequency in this bivariate distribution. From scrutiny of the points of tangency of these ellipses, he was led to identification of what we call the regression lines. A similar study relating heights of brothers, according to Stigler (1986), helped to free him from the temporal or intergenerational context of regression, and to reveal the mathematical symmetry of the two regressions, predicting either variable from the other. Noticing that these two regressions were in fact equal when the dimensions were rescaled in units of the probable error led him, finally, to a symmetric expression for the mutual linear relationship of two variables, in their coefficient of co-relation.⁷

Galton often made imaginative use of mechanical aids to exposition of his ideas. His system of weights and pulleys, essentially an analog computer, for predicting height of offspring from parents' heights (1889), is now little used, but his famous Quincunx led to one of his more important conceptual advances. He took the name from a Roman coin one of whose sides bore five spots in a pattern like that on a modern die. The device, with its quincunx array of pins through which shot is poured from a funnel, exhibits "in a very pretty way," as Galton said (p. 63), the generation of a bell curve. Experimenting with it led him to "the most important step in perhaps the single major breakthrough in statistics in the last half of the nineteenth century" (Stigler, 1986, p. 281). He intercepted the shot at an intermediate level, collecting it into compartments; provided the point of interruption were not too close to the top, the shot would fall into a bell-shaped distribution. If any single compartment were released, the result would be a "small" bell-shaped distribution at the bottom, centered under the compartment above. Since releasing all

⁶In an earlier day, it was fashionable to liken the shape of the curve to that of a constable's hat, or *chapeau de gendarme* (cf. Lottin, 1912/1969), but fashions change with the times. According to Bertrand (1889), the suggestion of a bell shape was due to a Professor Jauffret. It was Pearson (1894) who, endowing it with a profusion of dubious connotations, and thereby insuring acceptance of his label, called it the normal curve.

⁷Stigler (1986) suggests that Galton's initial spelling was an attempt to distinguish the concept from some earlier, nonstatistical uses of the word *correlation*, but Galton was soon speaking of an "index of correlation"; and Edgeworth's (1892) term "coefficient of correlation" was standard from the time he introduced it, for a formula essentially the same as that for which Pearson took credit.

compartments also generated an overall bell curve, Galton recognized that a mixture of such curves also resulted in a bell curve. This was essentially what his experiment with sweet peas showed. Its importance lay in the fact that this curve could now claim a broader rationale, and was no longer so narrowly tied to the simple binomial justification provided by Laplace and Quetelet.

For the purpose of the present history, Galton's other main contribution was simply to pass Quetelet's ideas on to Karl Pearson and to fund, out of his own personal fortune, a laboratory under Pearson for the study of biological variation and inheritance. By Pearson's account, he and Galton were brought together by W. F. R. Weldon, a biologist interested especially in heredity and evolution.⁸ Raphael Weldon saw in Galton's method a means of gathering new evidence for the Darwinian hypothesis, through statistical studies of organic correlation and the rate of change of species characteristics, rather than through embryology and morphogenesis. In 1891 Weldon was appointed Professor of Zoology at University College, London, where Pearson held a chair in Applied Mathematics; after having studied the French authors on probability (the only significant English contributions since Bayes, curiously, being those of Boole and Venn), he appealed to Pearson for further help. Pearson had been an engineer, but his interest in both statistics and evolution was stimulated by Weldon. He may also have become interested in probability through Isaac Todhunter; he took some of Todhunter's courses and later completed that author's *History of the Theory of Elasticity*; according to E. S. Pearson (1965), "The second volume of this, containing some 1300 pages, was almost entirely Pearson's contribution" (p. 3). It was, in any event, to these fields rather than engineering—to their marriage in the new science of biometrics—that Pearson devoted the rest of his life.

At Weldon's suggestion (K. Pearson, 1906), Galton, Weldon, and Pearson founded the journal *Biometrika* for their new science in 1901. Weldon died in 1906; Galton followed him in 1911, leaving funds for a chair in eugenics at the University College London, to which Pearson was appointed. The appointment seems natural in retrospect, given Pearson's urgent call for altering the relative fertility of rich and poor as a means of improving the race.⁹ Pearson (1904) argued by analogy that intellectual and moral characters must be inherited in the same way as physical characteristics. Systematic data were hard to come by, and research efforts were hampered by the lack of an intelligence scale, but Pearson devised his own 7-point rating scale and assumed normality for the resulting distribution of scores (Welch, 1968).

Pearson's early work in statistics was published in a series of long memoirs in the *Philosophical Transactions of the Royal Society of London* in the 1890s. In these he derived a whole family of probability distributions, comprising most of those in

⁸Stigler (1986), having examined the sequence of articles and correspondence closely, thinks Pearson's memory here was faulty, and that Francis Edgeworth had the more important role in connecting Galton and Pearson; but it makes little difference for the present history.

⁹It comes as no great surprise to learn that he was also a Germanophile; Haldane (1957) reports that it was evidently during a year in Germany as a young man that Pearson began spelling his name with a K instead of a C.

common use today in psychological research: The χ^2 is a special case of the gamma, the F distribution is a transformation of the beta, and the t distribution is of course closely related to both of these. The biometricians were concerned primarily with estimating parameters of natural populations and with making inferences about underlying processes from the shapes of distributions. Pearson thought that, just as Galton and Quetelet, reversing Laplace, had argued from a bell-shaped distribution to homogeneity of the underlying process, he could argue also from a well-fitted skew curve to an underlying homogeneous, if more complex, generative process. Weldon disagreed, contending that extreme cases, such as those resulting from breakage and regeneration of limbs, should be excluded and that, if they were, the remaining distribution could be fitted well enough by a normal curve.

The biometricians obviously could not study empirical populations exhaustively; but, like modern opinion pollsters, they took samples large enough to be representative of the population, within usefully narrow limits of sampling error—typically their samples comprised many hundreds of observations. The estimates they obtained were thus stochastically stable, and, partly as a result, the distinction between sample and population was not so salient as it is to us today; indeed, the very lack of distinction in notation between sample and population hindered the mathematical development of sampling theory.

Probabilistic aspects of the model appeared in the calculation of errors of estimate, which made reference necessarily, if only tacitly, to a context of repetitive random sampling. They also made tacit, and almost furtive, reference to inverse probability. Typically a mean would be reported, plus or minus the probable error, the probable error being that deviation which cuts off the most extreme 50% of a distribution. It is the point, in other words, which a random observation is as likely as not to exceed; in a normal distribution the probable error of estimate for a mean is 0.6745 times the standard error of the mean. As the method of interval estimation was usually understood at the time, the reporting of a mean plus or minus such limits implied that there was a 50% probability that the population mean lay within these bounds. To ascribe a probability distribution to the population mean, however, was to use the Bayes-Laplace concept of inverse probability that had been criticized by Boole and Venn. The philosophical uneasiness that accompanied the practice provided a gentle reminder of the need for a theoretically more coherent approach.

The biometric methodology per se seems, interestingly, to have yielded little for science. When Weldon found a bimodal distribution of head width in crabs, he assumed he was confronted with a mixture of two distinct populations, and sought to determine whether their relative proportions were changing owing to differential survival value. But Tankard (1984) observes that his method for the determination was classic rather than biometric: He put the two kinds of crabs in muddy water and showed that those with smaller heads had a higher survival rate, perhaps because of better filtration. Hogben's (1957) somewhat uncharitable appraisal, from the opposing camp of Bateson, is that Pearson retarded the progress of experimental genetics for half a generation. The historical importance of the movement, in any event, lay much more in its having served as the midwife to the emergence of the modern theory of statistical inference. Its contributions were to focus attention again on the

problem of inferring characteristics of populations from samples and on distributions which were now normal instead of binomial.

Three further developments were still necessary. One building block for the new theory of statistical inference was provided by the incipient practice of significance testing in the late nineteenth century. A second, whose relevance could not be foreseen, was the advent of so-called small-sample theory in agriculture in the early years of the twentieth century. The third was certain developments in the theory of probability, which will be covered in Chap. 6.

5.2 Precursors of Significance Testing

5.2.1 Arbuthnot's and Gavarret's Use of the Binomial

Actually, the idea of making a decision based on stochastic expectation became a possibility from the time the calculus of gambling came into being, and it was not long before that possibility was realized. Credit for the first “significance test” is given to John Arbuthnot (1710), who was also inventor, in a humorous pamphlet in 1712, of the archetypical Englishman John Bull (along with the Dutchman Nicholas Frog and the Frenchman Louis Baboon). Arbuthnot had written in the preface to his English translation of Huygens' *Rationiciis in Aleæ Ludo* in 1692:

I believe the Calculation of the Quantity of Probability might be . . . applied to a great many Events which are accidental, besides those of games; . . . all the Politicks in the World are nothing else but a kind of Analysis of the Quantity of Probability in casual Events, and a good Politician signifies no more, but one who is dexterous at such calculations. (quoted in Pearson, 1978, p. 140)

Arbuthnot appears clearly here to hold the same vision of the scope of probability as James Bernoulli, whose book was still 21 years from publication. Arbuthnot's contribution, also antedating de Moivre, was naturally what we would call nonparametric. As was mentioned in Chap. 2, Arbuthnot's test, unsurprisingly for its time, was an argument for the existence of God; but it was surprisingly modern in form, and it provides an interesting prototype.

Arbuthnot began, in the manner that is now standard, by framing a null hypothesis (albeit without calling it that): that there is no difference between the birth rates of males and females. He took from Graunt the records of christenings which had been kept since 1629 in London, providing him, up through 1710, with 82 years of data. If we make the dubious equation of the number of births with the number of christenings, then male births were in excess in all of the 82 years. Hence Arbuthnot, like a modern research worker, was quite convinced before he started that his null hypothesis was false. He acknowledged the variability of possible outcomes on the null hypothesis, then added, leading into the statement of his alternative, or theoretical, hypothesis:

But it is very improbable (if mere Chance govern'd) that they would never reach as far as the Extremities. But this Event is wisely prevented by the wise Oeconomy of Nature; and to judge of the wisdom of the Contrivance, we must observe that the external Accidents to which are Males subject (who must seek their Food with danger) do make a great havock of them, and that this loss exceeds far that of the other Sex, occasioned by Diseases incident to it, as Experience convinces us. To repair that loss, provident Nature, by the Disposal of its wise Creator, bring forth more Males than Females; and that in almost a constant proportion. (Arbuthnot, 1710, p. 188)

If male births are binomially distributed with $p = \frac{1}{2}$, then the probability that the number of male births will exceed the number of female births in a given year is also $\frac{1}{2}$. Since the concept of probability was not yet well established, Arbuthnot naturally cast his problem in terms of expectations, the language of the day, and proceeded to reason as follows:

Problem. A lays against B, that every Year there shall be born more Males than Females: To find A's Lot, or the Value of his Expectation.

It is evident from what has been said, that A's Lot for each Year is less than $\frac{1}{2}$ (but that the Argument may be stronger) let his Lot be equal to $\frac{1}{2}$ for one Year. If he undertakes to do the same thing 82 times running, his Lot will be $(\frac{1}{2})^{82}$, which will be found easily by the Table of Logarithms to be 1/4836000000000000000000000000. But if A wager with B, not only that the Number of Males shall exceed that of Females, every Year, but that this Excess shall happen in a constant Proportion, and the Difference lye within fix'd limits; and this not only for 82 Years, but for Ages of Ages, and not only at *London*, but all over the World; (which 'tis highly probable is Fact, and designed that every Male may have a Female of the same County and suitable Age) then A's chance will be near an infinitely small Quantity, at least less than any assignable Fraction. From whence it follows, that it is Art, not Chance, that governs. (pp. 188–189)

Arbuthnot's test was thus an example of the particular argument from design developed by Durham, Newton, and later Süssmilch, that stable statistical ratios are scientific evidence for the existence of God, who must intervene to maintain equilibrium in a world which, left to its own devices, would run wild.

By way of a discussion section, Arbuthnot concluded by arguing for a

Scholium. From hence it follows, that Polygamy is contrary to the Law of Nature and Justice, and to the Propagation of the Human Race; for where Males and Females are in equal number, if one Man takes Twenty Wives, Nineteen Men must live in Celibacy, which is repugnant to the Design of Nature; nor is it probable that Twenty Women will be so well impregnated by one Man as by Twenty. (p. 189)

Polygamy appears to have been a hot issue at the time, in much the same way that atheism was. Vigorous arguments against it were everywhere (e.g., in Nieuwentyt), yet there were precious few voices to be heard from the other side. I am aware only of Süssmilch's interestingly casual definition of marriage as "the union of a man with one or more women" (quoted in Pearson, 1978, p. 326). One infers that the temptation to both was strong.

Even as modern psychologists would stop short of inferring the existence of God from Arbuthnot's data, many might still see them as showing conclusively that "it was not chance that governed." In applications of the sign test, it is a common

fallacy even today to suppose that only the hypothesis $p = .5$ is compatible with chance. That error was very quickly pointed out by Nicholas Bernoulli in a letter to Montmort, published in the second edition of the latter's *Essay d'Analyse sur les Jeux de Hazards* (1713). Using his uncle's recently published "golden theorem" (J. Bernoulli, 1713), he showed that a binomial distribution with a probability of 18/35 for a male birth fit the data very closely.

Another feature of Arbuthnot's argument is worth noticing for future reference. As stated, it would make *any* outcome incompatible with his null hypothesis. For any specific outcome—for instance, 41 male and 41 female years in alternation—has probability $(\frac{1}{2})^{82}$ on this hypothesis. The principle to which Arbuthnott made implicit, if obvious, appeal was that the outcome of 82 male years was very much more likely under his alternative hypothesis. This principle is familiar to the last generation of psychologists raised on the Neyman-Pearson doctrine of Type I and Type II errors, but it has not always seemed self-evident: R. A. Fisher (1956/ 1973) was still arguing forcefully in his last book on statistical inference that no reference to alternative hypotheses was necessarily implied in a test of significance.

Arbuthnot did not call his procedure a significance test, nor did he claim for it the status of a new method of inference. It never attracted very much attention, apart from the Dutch mathematicians Nieuwentyt and 'sGravesande, and failed to start a trend.

Over a century later, Jules Gavarret (1840), a student of Poisson, applied the binomial distribution to medical statistics in a way that sounds about 50 years ahead of its time. He cited a report on the number of cholera patients admitted to Paris hospitals in a 27-week period, classified by days of the week (see Table 5.1). The authors of the report had argued on the basis of the "excess" admissions on Mondays among these working-class patients that drunkenness predisposed one to cholera! The χ^2 goodness-of-fit test being another 60 years away, Gavarret divided the week into halves in various ways (e.g., Sundays, Mondays, Wednesdays, and Thursdays against Tuesdays, Fridays, and Saturdays, the latter prorated by 4/3) to show that the difference between proportions was within what was to be expected on the basis of chance, using Poisson's criterion of $p = .9953$ (corresponding to $z = 2\sqrt{2}$), and

Table 5.1 Cholera admissions to Paris hospitals by days of the week

Days	Number of admissions
Sundays	1833
Mondays	2075
Tuesdays	1947
Wednesdays	1978
Thursdays	2004
Fridays	1971
Saturdays	1969

Note. From *Principes Généraux de Statistique Médicale* (p. 209) by J. Gavarret, 1840, Paris: Bechet Jeune et Labé

hence that there were no irregularities to be explained. His demonstration is remarkable for the early use of what we would call inferential statistics to settle disputes about data and for the advance specification of probability limits.

Gavarret wrote as though the day were not far off when the treatment of choice for various ailments might be determined by such statistical methods. He emphasized, as Bernard (1865/1952) would after him, that the cases must be similar or comparable, and he repeated at least three times through the book that the analyses must be based on a minimum of several hundred cases. Whether it was because the data were not forthcoming or because his ideas were too far ahead of his time, Gavarret's work attracted little notice, then or since; he is not mentioned in Stigler's (1986) encyclopedic history of nineteenth-century statistics.

Significance tests as we know them were not to become common, however, for another century, and when the practice finally caught on, it was by a more circuitous route, leading to so-called parametric tests.

5.2.2 Probabilistic Criteria for the Rejection of Discordant Observations in Astronomy

The normal distribution had been used to model errors of measurement in astronomy, and the modern use of the probability integral over the tails of the distribution as the criterion in significance testing evolved from probabilistic criteria for the rejection of outlying observations. Sets of measurements assumed to arise from a normal process often contain deviations large enough to be suspected of having resulted from extraneous factors; and it is reasonable, if the normal law were to be retained, that errors actually arising from extraneous causes should be deleted in striking the average and in calculating the probable error. (The alternative would have been to assume a single operative process and to adopt a law giving greater probability to large deviations.) In cases where the observer had specific grounds for questioning the validity of a particular observation (e.g., seeing the assistant who made the observation stumble away from the telescope), there was of course no controversy. The challenging situation arose where the magnitude of the observation itself was the only basis for doubt, and here the practice of rejecting observations was hotly debated. To many, it smacked of fraud. Daniel Bernoulli was taking a dim view of the procedure in 1777:

I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others. Nevertheless, I do not condemn in every case the principle of rejecting one or other of the observations, indeed I approve it, whenever in the course of observation an accident occurs which in itself raises an immediate scruple in the mind of the observer, before he has considered the event and compared it with the other observations. If there is no such reason for dissatisfaction I think each and every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken every care. (translated by C. G. Allen, in Kendall, 1961, p. 3)

Perhaps Bernoulli's attitude was considered to leave too much to the discretion of the observer in eliminating discordant observations. In any event Benjamin Peirce (father of Charles Sanders) proposed a strictly probabilistic criterion for rejection, based just on the probabilities of the observations on the assumption of a normal distribution. Peirce's (1852) procedure was a recursive scheme; the criterion for rejection was determined as follows. First the mean and variance of the n observations were calculated, and from them the joint probability of the whole set of n values, assuming a normal distribution. Then the most deviant observation was eliminated, and the mean and variance recalculated. The cut-off for rejection was set to minimize the probability of rejection; i.e., it was set at the value of the most deviant observation. The new probability of the whole set of n observations was then the product of this latter probability and the recalculated probability of the other $n - 1$ observations. If the probability for all n under this second scheme was smaller than the original calculation, all the observations were retained; if it was larger, the most deviant observation was dropped, and the cut-off was set at the second most deviant. The mean and variance were then computed from the remaining $n - 2$ observations, and their joint likelihood calculated; then this quantity was multiplied by the probability of having two observations in the (new) rejection region and compared with the previous probability for the whole set; and so on.

William Chauvenet (1863) soon introduced a simpler criterion for the probabilistic rejection of observations. Again using the sample data to estimate the mean and variance of the normal distribution represented by the observations, he multiplied the tail area by the total number of observations in the sample, to get the expected number of rejections; he set this quantity equal to $\frac{1}{2}$ and solved for the cut-off point. If any observations fell beyond this point, they were rejected, since on the normal law, only half an observation was expected beyond it. Chauvenet's criterion could also be applied recursively, like Peirce's, with a new mean and variance each time, until no further observations were rejected. He recommended it because it was simpler than Peirce's, yet gave similar results.

The probabilistic criteria for rejection of observations proposed by Peirce and Chauvenet came in for criticism just as the older, more subjective method had. George Biddle Airy (1856), Astronomer Royal at the Greenwich Observatory, essentially repeated Bernoulli's criticisms 80 years later: that the normal law explicitly holds that various improbable events should occur a certain proportion of the time, and we have simply observed one of them to occur.

In seeking for truth or accuracy in a matter of chance, we are bound to take all the evidence that we can find; every individual observation contributes its evidence; and if we have no *a priori* reason for preferring one observation to another while we are yet in the dark as to the result, and no cogent reason for supposing that unusual causes of error must have intervened in special observations, we are bound to admit all on the same terms, as giving equally valid evidence. (Airy, 1856, p. 137)

I have, not without surprise to myself, been led to think, that the whole theory is defective in its foundation, and illusory in its results; that no rule for the exclusion of observations can be obtained by any process founded purely upon a consideration of the discordance of those

observations; and that the only rule (imperfect as it may be) by which such exclusion can be justified, must be based upon considerations of a totally different kind. (p. 137)

Jeffreys (1961) addresses his criticisms more specifically to the particular criterion used. Chauvenet's criterion, he observes, gives an even chance of rejecting the most extreme observation even when the normal law holds true. The real problem, he contends, lies in the use of the tail integral, which includes nonoccurring results. Peirce (1852), who was originally responsible for the use of the tail integral, explained it by saying that because observations are rejected in the second system, they must be regarded, in the first system, not as being limited to their actual values, but only as exceeding the rejection criterion. Jeffreys, for his part, would use the ordinate of the normal distribution as the criterion, since the ordinate approaches zero as the magnitude of the deviation increases; and ordinate cut-offs can be selected to correspond to other criteria, like Peirce's or Chauvenet's, as desired. Thus, he says, when the question was whether a new parameter in the probability distribution (e.g., a nonzero mean) were needed to account for the observations,

It was habitually found that those up to about twice the standard error tended to diminish when the observations became more numerous or accurate, which was what would be expected if the differences represented only random error, but not what would be expected if they were estimates of a relevant new parameter. But this could be dealt with in a rough empirical way by taking twice the standard error as a criterion for possible genuineness and three times the standard error for definite acceptance. This would rest on a valid inductive inference from analogous cases, though not necessarily the best one. Now this would mean that the former limit would be drawn where the joint probability of the observations is e^{-2} of the value for the most probable result, supposing no difference present, and the latter at $e^{-4.5}$. This would depend on the probability of the actual observations and thus on the ordinate of the direct probability distribution, not on the integral. The ordinate does depend on the hypothesis and the observed value, and nothing else. (Jeffreys, 1961, pp. 386–387)

The trouble with the tail integral is the paradox of rejecting observations that may be normal on the ground that other observations *which have not occurred*—namely, any other results in the rejection region—*were not predicted* by the normal law (i.e., were given a low probability). We are today so accustomed to adding the phrase “...or results more extreme” that we do not notice anything odd in using nonoccurring results as part of our criterion for rejecting those that did occur. But Jeffreys regards the failure of the normal law to predict what did not happen more supportive of it than disconfirming.

Jeffreys offers here no insight as to *why* 2 standard errors should have habitually been found to be the cut-off with Chauvenet's criterion, perhaps because it seemed simply too obvious. The cut-off depends directly on the number of observations; since ± 2 standard errors cuts off roughly the outer 5% of the distribution, the number of observations must have been about 10 ($10 \times 0.05 = .05$ observations beyond the cut-off). And in fact this appears to have been a typical number of observations in astronomical studies. Some astronomical phenomena can be observed only annually, and mid-nineteenth-century astronomers were unlikely to have amassed more than a couple of decades' worth of observations. Other phenomena could be observed only for a few days in a row, and, from my rather cursory scan of *The*

Astronomical Journal for the time,¹⁰ seven observations seems to have been a common number. Merriman (1888) used an example with 13 observations to introduce Chauvenet's criterion. What I think we are seeing, here, in short, is the origin of the 5% convention—though it would take a few more interesting twists of history to conventionalize it.

It is also easy to get the impression that Chauvenet's criterion, despite its importance for future developments in other disciplines, was not a staple of late nineteenth-century work in astronomy. Doolittle (1888), in his textbook, appears to follow Airy in taking a dim view of the procedure:

As for the criteria for this purpose [rejection of discordant observations] hitherto proposed, probably the most that can be said in their favor is that their use insures a uniformity in the matter, thus leaving nothing to the individual caprice of the computer. (p. 32)

And Greene (1891), in his small textbook, makes no mention at all of the rejection of discordant observations.

5.2.3 *The Normal Model and Data in the Social Sciences*

Although Peirce's rejection scheme is recognizable as the original use of the tail integral, the normal model had meanwhile been transplanted to sociological data, and the idea was independently developing in that area of tests of discrepancy between model and data. The idea came slowly, however.

One early example was due to Robert Campbell (1859), whose article was inspired by Henry Thomas Buckle's *History of Civilization in England*, published 2 years earlier. Porter (1986) describes Buckle as “possibly the most enthusiastic and beyond doubt the most influential popularizer of Quetelet's ideas on statistical regularity” (p. 65). Buckle saw in Quetelet's work support for the ascendant doctrine of classical liberalism: The social laws that were revealed by statistical regularities demanded for their smooth operation an absence of forcible intervention by government. Campbell thought the stability of some of these series ought to be tested. He devised a multinomial test for suicide data—though, as Stigler (1986) notes, without drawing any conclusions. He described his procedure as follows:

We started with the hypothesis that we knew *nothing* about the laws regulating the phenomena and causing either uniformity or variety. Now this hypothesis was morally incorrect in this respect. For we knew *this* at starting, that such laws *may* exist; and if the degree of uniformity or the reverse in the real tables present a striking contrast to that which we should expect from the result of our hypothesis, we may fairly be led to the conclusion that such laws *do* exist, and that our tables justify us in looking for causes to explain their remarkable features. (Campbell, 1859, p. 366)

¹⁰ *The Astronomical Journal* suspended publication in 1861 for the American Civil War—but didn't resume publication until 1886. I assume it must not have taken that long for the smoke to clear, or for the staff to be repopulated, but I don't know anything more about the delay.

A more important and influential step was taken about the same time by Gustav Fechner, in his *Elemente der Psychophysik*. Though the book is best known for the Weber-Fechner law relating stimulus and sensation, Stigler (1986) sees its main contribution as methodological: “on experimental design, the most comprehensive treatment... before R. A. Fisher’s 1935 *Design of Experiments*” (p. 244). Of principal interest in the present context is Fechner’s use of Gaussian methods for modeling judgments of line length as having normally distributed random errors. Fechner did not refer to Quetelet, but gave credit for the suggestion of the normal model to Möbius—whom Stigler wryly describes as “clearly a many-sided mathematical talent” (p. 247). He deferred the question of testing his results to another book, which was never written.

His *Elemente der Psychophysik* was picked up in a secondhand Paris bookstore by Hermann Ebbinghaus, who used Fechner’s methodology in his famous book on memory in 1885. He calculated the probable error for all his averages, but to test the fit of his observations to a normal distribution relied on visual inspection (Stigler, 1986). Ebbinghaus’ work was influential in transmitting the normal model into modern psychology, but with none of these examples did the idea of testing discrepancies between model and observations really take hold, in psychological or sociological research.

In view of the apparent promise of that idea as a methodological tool, the slowness with which it caught on is puzzling to retrospect. The step that was evidently missing was that between frequency distributions and probability distributions. The concepts of sampling and sampling distributions were still in the future; and, if the “curve of frequency of error” were taken to represent the outcome of a random process, it was still conceived primarily as a device for insightful description rather than for resolving theoretical disputes. The first use of statistics for the latter purpose did not come until 1915, in an article by Yule and Greenwood on the effectiveness of inoculation, and they used Pearson’s χ^2 for a 2×2 contingency table rather than data from a normal distribution.

5.2.4 *The Pun on Significance*

The language of significance testing first appeared, at least “for good,” in the 1880s, in the work of Edgeworth and Venn. Among the contributors to the development of modern statistical inference, Francis Ysidro Edgeworth is one of the least well known to psychologists today. His role was primarily one of consolidator and catalyst; in at least the case of the correlation coefficient, credit for his specific innovation was claimed by someone else. Edgeworth’s early training, interestingly, was in classical literature; but he taught himself mathematics, and by 1885, when he was 40, his knowledge of the field, from Laplace to Fechner, was, according to Stigler (1986), unequaled in Britain. Edgeworth’s *Mathematical Psychics: An Essay on the Applications of Mathematics to the Moral Sciences* (see Chap. 1, Note 9), the rationalism of which strongly recalls Craig’s *Theologiae Christianae Principia*

Mathematica of two centuries earlier (Chap. 3), caught the favorable attention of Francis Galton, a distant cousin; and Stigler surmises that Galton may have been influential in the direction of subsequent development of Edgeworth's thought.

Although his terminology did not catch on, Edgeworth helped to clarify the distinction between applications of the normal model in astronomy and in the social sciences.

Observations and statistics agree in being quantities grouped about a Mean; they differ, in that the Mean of observations is real, of statistics is fictitious. The mean of observations is a cause, as it were the source from which diverging errors emanate. The mean of statistics is a description, a representative quantity put for a whole group, the best representative of the group, that quantity which, if we must in practice put one quantity for many, minimizes the error unavoidably attending such practice. Thus measurements by the reduction of which we ascertain a real time, number, distance are observations. Returns of prices, exports and imports, legitimate and illegitimate marriages or births and so forth, the averages of which constitute the premises of practical reasoning, are statistics. In short observations are different copies of one original; statistics are different originals affording one "generic portrait." Different measurements of the same man are observations; but measurements of different men, grouped round l'homme moyen, are *primâ facie* at least statistics. (Edgeworth, 1885a, pp. 139–140)

Like Galton's contemporaneous work, these concepts helped to free the normal model of its tie to errors of measurement, which had been prominent in Quetelet's thinking; they also helped prepare the way for explicit recognition of the concept of sampling distributions.

In another major paper the same year, Edgeworth gave examples from several different fields to illustrate what he called the science of means, which comprised two problems. One was estimation—deciding between the sample mean and median, for example; the other was "to find how far the difference between any proposed Means is accidental or indicative of a law" (1885b, p. 182). The second problem "investigates how far the difference between the average [actually observed] and the results usually obtained in similar experience where pure chance reigns is a significant difference; indicative of the working of a law other than chance, or merely accidental" (p. 182). As an example:

The rate of mortality among young farmers between the ages of 15 and 25, as based upon the observations of 65 deaths in the year, exceeds the rate of mortality in all other professions by 0.3 per cent. How far is such an extent of deviation based upon such a number of observations significant of a real difference in respect of healthiness between the conditions of young farmers and the rest of the industrial community? (p. 182)

This passage may be the first to talk about statistical significance. It reveals very clearly, in any case, the paronomastic origin of the term. A century ago, when scholars in all fields were fluent in Latin, the participial meaning of *significant* was much more salient than it is today; and that is clearly the meaning Edgeworth intended in introducing the term, as a synonym of *indicative*. In a succession of examples later in the article, he used the following language: the "difference in figures is indicative of a real difference" (p. 196); "the observed difference... certainly is not accidental, but indicates a law" (p. 196); "the observed difference is significant" (p. 196); "the observed difference is important" (p. 196); "the difference [is] very significant"

(p. 197). Here the pun becomes explicit and actually seems to emerge in the course of his writing. The qualifier *very* in the last example is of course applicable only to *significant* in the sense of *important*. That potential-laden double entendre was so attractively available that, had Edgeworth not picked it up, no doubt someone else soon would have; but we may still tentatively award him credit for this important step in the evolution of the concept of significance testing.¹¹ It is not easy to judge from this article alone whether Edgeworth's pun was intentional. He was certainly a man who was conscious of his use of language¹²; and, if he had wanted to ensure the immortality of his choice of term, he could have done no better. It seems less likely, however, that he would have foreseen how his term would catch on like wild-fire in the social sciences 70 years later than that he would simply have enjoyed the cleverness of the elision without making it more explicit.

Porter (1986) thinks that Edgeworth's direct influence was minimal because he "tended to define his own problems, often somewhat whimsically, rather than to proffer solutions to statistical questions under study at the time" (p. 269). I agree, but it hardly matters. Edgeworth did pass his ideas and terminology on to Venn, who promptly incorporated them into the third edition of *The Logic of Chance* (1888), particularly in a new chapter entitled "The Theory of the Average as a Means of Approximation to the Truth"; and Venn's wide audience included Ronald Fisher, who was born in 1890. Venn (1889) also spoke at this early date of statistical significance in anthropometric research—though Hogben (1957) is scornful of this particular application. "All this [added] up to little in terms of the world's work," he contends, "because anthropometry had then (as now) scanty, if any, practical value other than for manufacturers of ready-made wearing apparel or of school furniture" (pp. 325–326). However, "Under the impact of the Darwinian doctrine in the setting of Galton's racialist creed and of the controversy over slavery in America, anthropometry [had become] a fashionable academic playground" (p. 325).

Galton had collected anthropometric data on 1095 Cambridge students—height, weight, grip strength, visual acuity, and several other measures—and passed them along to Venn for analysis. Venn's particular interest lay in the relation between

¹¹ It is worth noting, nevertheless, that the neglected Gavarret (1840) had used suggestively similar language 45 years earlier.

A difference established between the reports furnished by two long series of observations, has no real meaning [*signification*] except from the moment when it surpasses a certain *limit* which depends on the number of facts observed. (p. 98)

The variations observed were too small to have the meaning [*signification*] attributed to them. (p. 222)

¹² Annoyingly so, one infers. Keynes, capable himself of a well-turned phrase, complained of Edgeworth's writing that "Quotations from the Greek tread on the heels of the differential calculus, and the philistine reader can scarcely tell whether it is a line of Homer or a mathematical abstraction which is in course of integration" (quoted in Barbé, 2010, p. 82). And Robert Graves tells the story of Edgeworth's having met T. E. Lawrence at the station on his return from London. Edgeworth inquired: "'Was it very caliginous in the metropolis?' 'Somewhat caliginous, but not altogether inspissated,' Lawrence replied gravely" (quoted in Barbé, 2010, p. 237).

physical and intellectual characteristics of what he acknowledged was a fairly homogeneous sample of “the upper professional and gentle classes” (p. 143). For the intellectual rating, he had the tutors at the various colleges classify the men into three groups. “By A is meant a first-class man, in any Tripos examination, or one who is a scholar in his college. By B is meant all the remaining ‘honour men’ and by C those who may be called ‘pool-men,’ *i.e.*, candidates for the ordinary degree” (Venn, 1889, p. 143). Venn wrote:

When we are dealing with statistics, we ought to be able not merely to say vaguely that the difference does or does not seem significant to us, but we ought to have some test as to what difference would be significant. For this purpose appeal must be made to the theory of Probability. Suppose that we have a large number of measures of any kind, grouping themselves about their mean in the way familiar to every statistician, their degree of dispersion about this mean being assigned by the determination of their “probable error.” (p. 147)

One of the interesting findings was that the A group had worse eyesight than the Cs, a difference of about 1.3 inches (the measure was the distance at which “the ordinary little shilling prayerbook could be read” (p. 141). Venn points out that the difference was numerically greater among just the 24-year-old men in each group; but, whereas the whole-group difference was “highly significant,” the difference among the 24-year-olds was not at all, because of the much smaller sample. These calculations, he said,

inform us which of the differences in the above tables are permanent and significant, in the sense that we may be tolerably confident that if we took another similar batch we should find a similar difference; and which of them are merely transient and insignificant, in the sense that another similar batch is about as likely as not to reverse the conclusion we have obtained. (pp. 147–148)

The language of Edgeworth and Venn sounds familiar enough to us a century later that it is somewhat surprising that statistical inference as we know it would not become common for another 60 years or so. The reason has to do with the odd role of small-sample theory.

5.2.5 *Small Datasets in Agricultural Research*

A key figure in this development was William Sealy Gosset, though his direct contribution, like Edgeworth’s, was modest. Gosset was a chemist for the Guinness brewers in Dublin, and he had become heavily involved in statistical work as he was helping the firm analyze the results of experiments conducted to improve the quality of their barley. Comparative agricultural experimentation was nothing new,¹³ but its

¹³ It goes back at least to Bacon’s (1627) experiments on the acceleration of germination. Bacon steeped wheat seeds for 12 hours in a variety of different concoctions—cow, horse, and pigeon dung, powdered chalk, soot, and ashes, each mixed with four parts water; bay salt, mixed with about eight parts water; human urine; and claret, Malmsey, and spirit of wine. He used two controls, both unsteeped; one batch of seeds was watered, the other was not. The wheat steeped in urine came up the “Highest, Thickest, and most Lustie” (1627, p. 110); the urine was followed in

growth was greatly spurred by food shortages following the potato famines and, of course, World War I. Analyses of agricultural research data had not theretofore made use of probability theory. Gosset was familiar with the work of the biometrics group, and it somehow occurred to him that his data from field trials might be regarded as constituting a sample. Eighty years later, of course, we are accustomed to thinking of any set of numbers as a sample, regardless of whether any act of sampling took place, but Gosset's view represented a substantial leap. Thinking of his data in that way made the elaborate descriptive theory of the biometricians available, but that was still of very limited help. The problem lay in the variability of estimates of the population standard deviation based on small samples. Experiments could be repeated, as a means of assessing the sampling variability of the standard deviation of the mean—the brewers had a legitimate interest in the average yield—but, as Gosset observed, “Owing however to secular change, the value obtained [is] nearly always too low, successive experiments being positively correlated” (Student, 1908, p. 2).

Gosset was not the first to notice the problem, but he was the first to worry about it, just because he was doing applied rather than basic research, and his results had commercial value (MacKenzie, 1981). His contribution was to guess correctly the distribution of the variance of a sample from a normal population, to show that the correlation between the sample variance and the mean was zero, thus to assume (correctly, in this case, as it happens) their mutual independence and to derive what Fisher was later to call the *t* distribution. Gosset's derivation was his own, and the only one known to Englishmen at the time, though Pearson later found the same result in a German paper by Helmert in 1876; that claim of priority was honored until recently, when a Russian (Sheynin, 1971) found that Abbé, also writing in German, had given the distribution of the variance in 1863. It was characteristic of Gosset that he derived several important results afresh, without much éclat; a highly talented but unassuming man, he was, as Yule said on meeting him, “a very pleasant chap, not at all the autocrat of the *t*-table” (Kendall, 1952, p. 159). His use of a pseudonym was not, however, a matter of modesty, but of company policy—though it is not clear whether Guinness prohibited their employees from making any sort of

effectiveness by the dungs, chalk, soot, ashes, salt, the unwatered control, the watered control, and the claret. The Malmsley and spirits were not only inferior to the control, but did not germinate at all. Bacon advised that:

This *Experiment* would be tried in other *Graines, Seeds, and Kernells*: For it may be some *Steeping* will agree best with some *Seeds*. It would be tried also with *Roots* steeped as before, but for *longer time*. It would be tried also in *Severall Seasons of the Yeare*. . . . (p. 110)

His publisher, William Rawley, acknowledged in his foreword that he had heard “his Lordship discourse, that Men (no doubt) will thinke many of the *Experiments* contained in this Collection, to bee Vulgar and Triviall; Meane and Sordid; Curious and Fruitlesse”; but he offered the Baconian defense:

As for the *Vulgarnes* of them, true Axiomes must be drawne from plaine Experience, and not from the doubtful; And his Lordships course is, to make Wonders Plaine, and not Plaine things Wonders; and that Experience likewise must be broken and grinded, and not whole, or as it groweth.

public statement or whether they were trying to keep from their competitors the fact that they employed trained research workers; the ban was in any event lifted shortly after Student's death, just before World War II (Tankard, 1984).

Gosset illustrated his *t* technique (which he called *z*) with several examples, one of them an experiment recently reported by Cushny and Peebles (1905).¹⁴ These authors had administered dextrohyoscyamine hydrobromide to 10 patients in a study of its effectiveness as a soporific. They found an average increase in sleep of three-quarters of an hour, with a range from -1.6 to +3.7 hours. Gosset's conclusions were the same as theirs. He worked out a probability of .887, or roughly 90%, for getting a mean less than 0.75 on the assumption that the drug had no effect; in the terminology of the future, he had obtained a one-tailed significance level of a little more than 10%. He interpreted the result as providing mild, but not striking, support for the presence of an effect, as Cushny and Peebles had without benefit of the *t* test.

It is worth noting that Gosset interpreted his *t* technique in terms of inverse probability, as was usual for the time. He sought "the probability that [D-hyoscyamine hydrobromide] will on the average give increase of sleep; i.e., what is the chance that the mean of the population of which the [10] experiments are a sample is positive" (1908, p. 20). His concluding statement was cast very much as a contemporary Bayesian analysis: "The odds are .887 to .113 that the mean is positive" (p. 21). His idea here was in fact very much like Fisher's notion of fiducial probability some two decades later (Chap. 7): If there is an 11% probability of the sample mean lying more than 0.75 above the population mean of 0, there is ipso facto an 11% probability of the population mean lying more than 0.75 below the sample mean of 0.75 (i.e., being less than 0). Gosset unwittingly shifted from a sampling distribution of means about a population mean to some sort of distribution of the population mean about the sample mean.

We might expect Student's *t*, formally allowing the extension of statistical theory into work with small samples, to have been seized upon as a significant advance. Remarkably, to retrospection, it was almost totally neglected for 30 years. The problem was very simply that, so long as investigators saw themselves as inquiring into actual populations, and estimation of their parameters, small-sample theory had very little to offer. Even knowing the sampling distribution of the mean was of limited value when the variance was so large; estimates based on 10 or 20 or 30 observations were so highly variable, so imprecise, as to be nearly worthless. The fact that probabilities could be attached to such estimates remained a mathematical curiosity. Pearson and his biometrics group, though friends and correspondents of Gosset's, were skeptical of small-sample results, and did little to encourage use of the *t*. Mises (1928/ 1957), on the Continent, went so far as to reject small-sample theory altogether. Referring to Gosset's analysis of the Cushny and Peebles data, with a probability of 90% for a lower sample mean on the hypothesis of no effect, he wrote:

¹⁴ Gosset did not give the complete reference, but only the year, which he listed incorrectly as 1904.

I do not think that any sensible doctor will have much confidence in this figure of 90%.

According to our way of thinking, if we have ten observations whose results oscillate between plus 3.7 and minus 1.6, we cannot draw any conclusions unless we include some a priori information, namely, some knowledge concerning the drug, gained independently of and in addition to our ten experiments. If it is impossible or too difficult to find a numerical expression for such a priori knowledge, we have no other recourse but to extend our sequence of observations to many hundreds or thousands of cases. (p. 159)

He went on to say: “As a matter of fact, it almost seems that the heyday of the small-sample theory, after a rather short duration, is already past. There is little reference to it nowadays” (p. 159). Needless to say, Mises’ underestimation of the impact of R. A. Fisher was gross.

References

- Airy, G. B. (1856). Letter to the editor. *Astronomical Journal*, 4, 137–138.
- Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London*, 27, 186–190.
- Bacon, F. (1627). *Sylva sylvarum, or, a naturall historie in ten centuries*. London, UK: William Rawley.
- Barbé, L. (2010). *Francis Ysidro Edgeworth: A portrait with family and friends*. Northampton, MA: Edward Elgar.
- Bernard, C. (1952). *Introduction à l'étude de la médecine expérimentale* [Introduction to the study of experimental medicine]. Paris, France: Flammarion. (Original work published 1865).
- Bernoulli, D. (1777). Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimilla inductione inde formanda [The most probable choice between several discrepant observations and the formation therefrom of the most likely induction]. *Acta Academiae Scientiarum Imperialis Petropolitanae*, I, 3–23. (Reprinted in Kendall, 1961, pp. 3–13; C. G. Allen, trans.).
- Bernoulli, J. (1713). *Ars conjectandi* [The art of conjecturing]. Basel: Thurnisius.
- Bertrand, J. (1889). *Calcul des probabilités* [Calculus of probabilities]. Paris, France: Gauthier-Villars.
- Campbell, R. (1859). On a test for ascertaining whether an observed degree of uniformity, or the reverse, in tables of statistics is to be looked upon as *remarkable*. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (4th series), 18, 359–368.
- Chauvenet, W. (1863). *A manual of spherical and practical astronomy. Vol. 2. Theory and use of astronomical instruments* (5th ed.). Philadelphia: Lippincott.
- Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités* [Exposition of the theory of chances and probabilities]. Paris, France: Hachette.
- Cushny, A. R., & Peebles, A. R. (1905). The actions of optical isomers. II. Hyoscines. *Journal of Physiology*, 32, 501–510.
- Daston, L. J. (1987). Rational individuals versus laws of society: From probability to statistics. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 295–304). Cambridge, MA: MIT Press.
- David, F. N. (1962). *Games, gods and gambling*. New York: Hafner.
- de Moivre, A. (1718). *The doctrine of chances*. London, UK: W. Pearson. (2nd ed., 1738; 3rd ed., 1756).

- Doolittle, C. L. (1888). *A treatise on practical astronomy, as applied to geodesy and navigation* (2nd ed.). New York: Wiley.
- Edgeworth, F. Y. (1885a). Methods of statistics. *Jubilee Volume of the Statistical Society* (pp. 181–217).
- Edgeworth, F. Y. (1885b). Observations and statistics. *Transactions of the Cambridge Philosophical Society*, 14, 138–169.
- Edgeworth, F. Y. (1892). Correlated averages. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (5th series), 34, 190–204.
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). New York: Hafner. (1st ed., 1956).
- Galton, F. (1885). The measure of fidget. *Nature*, 32, 174–175.
- Galton, F. (1889). *Natural inheritance*. London, UK: Macmillan.
- Galton, F. (1905). Number of strokes of the brush in a picture. *Nature*, 72, 198.
- Gavarret, J. (1840). *Principes généraux de statistique médicale, ou développement des règles qui doivent présider à son employ* [General principles of medical statistics, or the development of rules which should govern their use]. Paris, France: Bechet Jeune et Labé.
- Greene, D. (1891). *An introduction to spherical and practical astronomy*. Boston: Ginn.
- Haldane, J. B. S. (1957). Karl Pearson, 1857–1937. *Biometrika*, 44, 303–313.
- Hankins, F. H. (1908). Adolphe Quetelet as statistician. *Studies in History, Economics and Public Law*, 31, 443–576.
- Hirschman, A. O. (1977). *The passions and the interests: Political arguments for capitalism before its triumph*. Princeton: Princeton University Press.
- Hogben, L. (1957). *Statistical theory: The relationship of probability, credibility and error*. New York: Norton.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Clarendon Press. (1st ed., 1939).
- Kendall, M. G. (1952). George Udny Yule, 1871–1951. *Journal of the Royal Statistical Society*, 115, 156–161.
- Kendall, M. G. (1961). Studies in the history of probability and statistics. XI. Daniel Bernoulli on maximum likelihood. *Biometrika*, 48, 1–18.
- Keynes, J. M. (1973). *A treatise on probability*. New York: St. Martin's Press. (Original work published 1921).
- Lottin, J. (1912). *Quetelet: Statisticien et sociologue* [Quetelet: Statistician and sociologist]. Louvain: Institut Supérieur de Philosophie. (Reprinted by Burt Franklin, New York, 1969).
- Mackenzie, D. (1981). *Statistics in Britain, 1865–1930: The social construction of scientific knowledge*. Edinburgh: Edinburgh University Press.
- Merriman, M. (1888). *A textbook on the method of least squares* (3rd ed.). New York: Wiley.
- Mises, R. v. (1957). *Probability, statistics and truth* (2nd English ed.). New York: Macmillan. (Original work published 1928).
- Montmort, P. R. de (1713). *Essay d'analyse sur les jeux de hazard* [Analytical essay on games of chance] (2nd ed.). Paris, France: Quillau.
- Pearson, E. S. (1965). Some incidents in the early history of biometry and statistics, 1890–94. *Biometrika*, 52, 3–18.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London (Series A)*, 185(Part 1), 71–110.
- Pearson, K. (1904). On the laws of inheritance in man. II. On the inheritance of mental and moral characters in man, and its comparison with the inheritance of physical characters. *Biometrika*, 3, 131–190.
- Pearson, K. (1906). Walter Frank Raphael Weldon, 1860–1906. *Biometrika*, 5, 1–52.
- Pearson, K. (1924). Historical note on the origin of the normal curve of errors. *Biometrika*, 16, 402–404.
- Pearson, K. (1978). *The history of statistics in the 17th and 18th centuries, against the changing background of intellectual, scientific, and religious thought* (E. S. Pearson ed.). London, UK: Griffin.

- Peirce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal*, 2, 161–163.
- Porter, T. M. (1986). *The rise of statistical thinking 1820–1900*. Princeton: Princeton University Press.
- Quetelet, A. (1828). *Instructions populaires sur le calcul des probabilités* [Popular lessons on the calculus of probabilities]. Brussels, Belgium: Tarlier et Hayez.
- Quetelet, A. (1835). *Sur l'homme et la développement de ses facultés, ou Essai de physique sociale*. Paris: Bachelier. (Translated as *A treatise on man and the development of his faculties* by R. Knox. Edinburgh: William and Robert Chambers, 1842. Facsimile reproduction, Scholars' Facsimiles & Reprints, Gainesville, FL, 1969).
- Quetelet, A. (1846). *Lettres sur la théorie des probabilités* [Letters on the theory of probabilities]. Brussels: Hayez.
- Quetelet, A. (1848). *Du système social et des lois qui le régissent* [On the social system and the laws which govern it]. Paris, France: Guillaumin.
- Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.
- Sheynin, O. B. (1971). Studies in the history of probability and statistics. XXV. On the history of some statistical laws of distribution. *Biometrika*, 58, 234–236.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
- Student. (1908). The probable error of a mean. *Biometrika*, 6, 1–25.
- Tankard, J. W., Jr. (1984). *The statistical pioneers*. Cambridge, MA: Schenkman.
- Venn, J. (1888). *The logic of chance* (3rd ed.). London, UK: Macmillan. (1st ed., 1866).
- Venn, J. (1889). Cambridge anthropometry. *Journal of the Royal Anthropological Institute*, 18, 140–154.
- Welch, B. L. (1968). *The biometricians and the development of statistics*. Leeds, England: Leeds University Press.

Chapter 6

The Frequency Theory of Probability



The prevailing philosophical climate of the century from 1850 to 1950 is fascinating for the way in which it managed to keep alive the “hostile twins” of rationalism and empiricism from the seventeenth century.

With respect to theology, the impact of the Scientific Revolution was similar in form to that of the Renaissance a century or so earlier. As was noted in Chap. 2, the effect of the Renaissance, at least according to the standard account of Hacking (1975) and others, was to shift the source of knowledge from the books of men to the “great book of nature,” but the religious orientation was preserved: It was still God’s word that was being read. Although the Church felt justifiably threatened again by the advent of the Scientific Revolution, in fact the result was anything but a wholesale conversion to atheism. Instead, the overwhelming response of the natural philosophers of the seventeenth and eighteenth centuries was to try to ground religion in science. These theological debates, which persisted well into the nineteenth century, inevitably carried some of the idle character of the scholastic debates which the Scientific Revolution had tried to throw off—idle in that the issues could not well be resolved by the accredited methods of science. Two hundred years after the charter of the Royal Society constrained its members from “meddling with Divinity [or] Metaphysics,” scientists were finally to respond to the call of Comte (1830–1842/1864) for leaving metaphysics behind and sticking with positive data. It is tempting to suppose that the theological arguments were by that point simply running out of steam, but old systems of belief, no matter how little they may have going for them, don’t fade away so easily; dogma-eat-dogma world that it is, they hang on until they are coopted or pushed aside by something else.

It was not until 1834 that the term *scientist* was first used, by Whewell, at least in its modern meaning (Yeo, 1986); prior to that time, it would presumably have been too closely linked with the medieval *scientia*, which connoted certainty rather than the probabilistic generalizations to which the new science aspired. It may be that that new identity marked a transition to maturity, a greater security and autonomy in the field, growing out of accumulating theoretical and technological achievements.

What was developing around midcentury, in any event, was a programmatic rejection of dogma, of metaphysics, of rationalism that was wholly dogmatic and rationalist in character. That paradox was to be, if anything, even more pronounced in the twentieth-century philosophy of logical positivism than it was in the positivism of Comte himself, who tended to take a more developmental perspective.

Each new “generation,” as it were, from the Renaissance to the Scientific Revolution to the modern positivist program, thought that it had thrown off the yoke of dogma, finally, when in actuality it had only found another that was so comfortable it wasn’t noticed until a generation or so later. It hardly needs to be said that science has never thought of itself as a secular religion, and would not easily recognize itself in such a description now. Being so ruthlessly empiricist in its content, it has been largely unconscious of the equally unyielding, hypertrophied rationalism of its form. Nevertheless, its having taken over some of the functions of religion—providing an alternative way of relating to the world, a framework, however austere, for meaning, as well as the means of achieving a sense of security and control—may help to explain why specifically religious doctrine and mythology began a gradual decline which continues to this day.

In a philosophical climate sensitized to apriorism, when the more outrageous results of probability theory began to arouse uneasiness in the mid-nineteenth century, it was natural for critical scrutiny to fall on the principle of indifference. That postulate seemed impossible to dispense with entirely; neither was it easily subjected to empirical test, on pain of circularity: Any such test ultimately appealed to probability. It was also coming to be recognized that Laplace’s definition was circular: The “equally possible” cases in the denominator refer back, in only the most thinly veiled way, to probability.

Bernoulli’s Theorem appeared to offer the way out. It proved that the empirical relative frequency converged, in all probability, to that *a priori p*, unknowable except through the principle of indifference. Hence the latter might be dispensed with, and probability defined instead merely in terms of relative frequency. Ultimately that approach would lead to the attempt to repudiate the epistemic meaning altogether; but, in the nineteenth century, that traditional meaning was still too strong, and the early frequency theorists would struggle more openly with their ambivalence.

The introductory discussion of probability in Chap. 1 pointed out that there are two senses in which we might speak of relative frequencies: the “relative frequency” of an ace, for example, among the six sides of a die, and the relative frequency of an ace in a long series of tosses of the die. Bernoulli’s Theorem established a mathematical relation between the two, but they are, even on that very account, distinct. Many writers, even up through Neyman (see Chap. 7), have nevertheless confused them, sometimes with the effect of adding plausibility to the frequency theory. If, for example, we have a bag with r red and w white marbles, then most writers agree with Nagel (1945) in insisting that the proportion of red marbles in the bag is not identical to, and not to be confused with, the long-run relative frequency with which a red marble would be drawn. But Donald Williams (1945), a modern proponent of the classical, Laplacean theory, argues explicitly for the confusion others, like Gottinger (1974), may have committed unwittingly: “There is one very intimate

connection between the marbles in the bag and the marbles drawn which Mr. Nagel and his colleagues strangely overlook, but which is relevant to his quandary here: they are the same marbles” (p. 621). The frequentists are logically committed to maintaining the distinction, however, and the challenge to them is to dispense with classical probabilities altogether.

6.1 The Principle of Indifference

Before commencing the historical development of the frequency theory, it will be useful to consider in what respects the principle of indifference, and along with it the classical definition of probability, offered itself as a target. Daston (1988) suggests that the principle had its origin in the concept of equitable contracts and that the symmetry of gaming devices seemed self-evidently to support the equiprobability assumption. In Hume, for whom mathematical, philosophical, and psychological aspects of probability were very close (Chap. 3), the mathematical or philosophical principle of indifference reflected primitive psychological states undisturbed by passions or other distorting influences. When probability replaced expectation as the fundamental concept, the principle began to look more circular; and when psychology came to be distrusted, by the French authors, as a basis for probability, the principle was undermined still further. Many applications, nevertheless, seemed straightforward, and the results, however technically difficult, noncontroversial.

The famous Rule of Succession was an exception. The rule was derived by a particular application of the principle of indifference, to probabilities themselves: Confronted with an unknown probability, in the absence of knowledge favoring one value over another, we can assume all possible values between 0 and 1 equally likely. The great power of the Law of Succession attracted critical scrutiny, which focused on the principle of indifference. Though it represented but one particular application, it had the effect, fairly or not, of discrediting the principle itself.

Its demise, as an accredited basis for probability assessment, was insured with further interesting paradoxes and dilemmas in the late nineteenth and early twentieth centuries. One example was already mentioned in Chap. 4: the ambiguity in whether different ratios of black to white tickets in the urn should be considered equally likely, or different constitutions of the urn. A better known example can be presented as follows. Suppose that the last stop Amtrak makes before it gets here is 120 miles away, and that the trip never takes less than 2 hours and never more than 4. (Suppose.) Without any knowledge of its track record, I might suppose all possible values for the duration of the trip between 2 and 4 hours to be equally likely. In particular, the trip should be as likely to take less than 3 hours as to take more than 3 hours. If I am ignorant of travel times, however, I am equally in the dark as to average speed, except that it will range between 30 and 60 mph. But a time of 3 hours corresponds to a speed of 40, not 45 mph; so I cannot have a uniform distribution of probability for speed if I have one for time; yet my ignorance of the two variables is precisely equal.

The contradiction can perhaps be handled by (a) observing that the reciprocal is a nonlinear transformation and inferring that the principle of indifference holds only over linear transformations, and hence that a problem must specify the scale to be used; (b) blaming the paradox on the continuum, as was done with Zeno; or (c) both (a) and (b). But both of these are restrictions whose need, if it is real, had not been noticed before.

Another problem involving the continuum is also worth notice: the famous paradox of Bertrand (1889). A chord is drawn at random in a circle; what is the probability that it will be shorter than the side of the inscribed equilateral triangle? Bertrand gave three solutions:

1. It makes no difference at what point one end of the chord lies, so let one end be fixed and choose the direction of the chord at random. Then the chord will be shorter than the side of the inscribed equilateral triangle unless it falls within the central 60° arc of the total 180° range; hence the probability in question is $\frac{2}{3}$.
2. It makes no difference if we fix the direction of the chord and choose at random a place for it to cut the perpendicular diagonal. But to be shorter than the side of the inscribed equilateral triangle, the chord must not cut the diameter within half a radius of the center. That leaves half the diameter available, so the probability is $\frac{1}{2}$.
3. It makes no difference where we choose the midpoint of the chord; we may fix the midpoint and choose a direction at random. The chord will be shorter than the side of the inscribed equilateral triangle so long as the midpoint does not fall within half a radius of the center. The area within half a radius of the center is $\frac{1}{4}$ the area of the whole circle, leaving a probability of $\frac{3}{4}$ that the midpoint will be chosen outside.

The paradox can be handled, again, by disallowing infinite sets, or by rejecting the question as ill-defined, at least until a method of drawing the chord is specified; Bertrand himself argued for both. But the qualifications attached to the principle of indifference are increased again.

The most compelling example happens to involve discrete probabilities, and concerns, moreover, an application of substantial importance. Suppose we have r indistinguishable balls to be distributed into n cells. Then, if the assignment of a ball to one of the n cells is independent of the other assignments, it would be natural for a modern student of probability to assume each of the n^r arrangements of balls into cells equally likely. That assumption at least seemed obvious to Maxwell and Boltzmann when they set about to give a statistical description of the distribution of particles in regions of space. It turns out, however, that Maxwell-Boltzmann statistics (*statistics* here used in a sense peculiar to mechanics) does not apply to any known particles. The distribution of certain particles—photons, among others—can be described if we assume equal probabilities for distinguishable arrangements only, of which there are $\binom{n+r-1}{r}$ (cf., e.g., Feller, 1957). This scheme is known as Bose-Einstein statistics. Finally, the distribution of some particles—including

protons, neutrons, and electrons—requires Fermi-Dirac statistics. This statistics is the same as Bose-Einstein statistics, with the added condition that no more than one particle may occupy a given region; hence the $\binom{n}{r}$ possibilities are considered equally likely.

The success or failure of these three models for different types of particles could scarcely have been anticipated, and it looks very much as if the identification of equiprobable alternatives involves an ineradicable empirical component. Logic alone cannot dictate the appropriate equiprobable sets for the behavior of photons versus protons. If it took so long for the point to become obvious, it was because the paradigmatic applications of the principle of indifference—namely, games of chance—involved materials and procedures with which we are quite familiar. But Mises (1928/1957) argues that even here the assumptions of common sense tacitly include a large fund of information or judgments. Consider appeals to the symmetry of a die, for example.

One part of the ivory of which the die is made was certainly nearer to the tip of the tusk than some other part; consequently, the identity of behaviour of all parts is no longer a logical necessity. This identity of behaviour follows in fact from experience, which shows that the original position of the ivory on the animal does not influence its properties in this respect. . . .

If a supporter of the a priori concept of probability is pressed to explain what he understands by “complete homogeneity,” he finally merely requires that the centre of gravity of the cube should coincide with its geometrical centre. If he knows enough mechanics, he adds that the twelve moments of inertia about its twelve edges must all be equal. No one will any longer maintain that it is evident apriori that just these conditions are necessary and sufficient for the “equal possibility” of the six sides of the die, and that no further conditions, such as conditions involving moments of higher order, need be considered. In fact, this formulation contains a number of results taken from the mechanics of rigid bodies, a science likewise based on experience.¹ (Mises, 1928/1957, pp. 72–73)

In recent years, Mellor (1971), in connection with the propensity theory of probability (see Chap. 10), has attempted to formulate the principle of indifference in a positive way and to bestow on it some heuristic value. Instead of assigning equal probabilities through ignorance, he says, we are entitled to treat alternatives as equiprobable if, but only if, we know that the available evidence provides a reason for not preferring any one to any other. Evidence, on this formulation, can as well provide a reason for a preference as for lack of one. There is, moreover, no necessity to partition the field into ultimately indivisible, equiprobable alternatives, as Keynes (1921/1973) and Kneale (1949) after him tried to do; those attempts foundered particularly on the simple example of a biased coin. Mellor relies instead, for his probability assignments, on a principle he takes over from N. R. Campbell and G. Schlesinger: the principle of *connectivity*, which asserts basically that two

¹The point is easily overlooked in many other contexts as well. Inhelder and Piaget (1958) presumed themselves to be posing logical tasks for children in asking them, for example, about the behavior of billiard balls on a table, and missed the significance of the child’s understanding of the physical apparatus (see Falmagne, 1975).

physical systems never differ only in a single respect; as a corollary, that any theoretical entity that is detectable in just one way is unacceptably ad hoc. The principle is intended to be more heuristic than substantive. Mellor takes it to supply his needed positive evidence for assignment of probabilities: An asymmetry in the propensity of a coin to land heads or tails must be connected to some other asymmetry in its properties; conversely, the absence of such asymmetry is a (positive) ground for asserting symmetry in the propensity.

Mellor's formulation (or others like it) as heuristic allows us to treat any given probability assignment on that basis as "functionally a priori," to use the serviceable term of Beth and Piaget (1961). As with other theoretical models, we assume its truth provisionally as a means of testing its adequacy. In cases of poor fit, like subatomic particles, the model is discarded and a better one sought. Such a conception could evidently have come about only in the late twentieth century, when philosophy had begun to free itself, through the work particularly of Quine (1951) and Peikoff (1967/1979), from the straitjacket of the analytic-synthetic dichotomy. The nineteenth century, however, was securely lodged in that split, newly and cogently articulated by Kant. It was necessary that the principle of indifference either provide an absolute logical basis for probability assignments or that it be derivable from experience; there was no room for anything to move back and forth.

6.2 The Frequency Theorists

Keynes (1921/1973) accords actual priority for a frequency conception of probability to Leslie Ellis and to Cournot—to the former, for a paper read in 1842 (but not published until 1849), and to the latter for his *Exposition de la Théorie des Chances et des Probabilités* (1843); but John Venn (of diagram fame), whom they beat by 20-odd years, developed his theory more elaborately and so is usually given the roses. It is actually only in the most marginal sense that Ellis and Cournot may be identified as originators of the frequency theory. Certainly neither of them was setting out to purify the concept of probability of subjectivist connotations in the way that full-blown frequency theorists would. Cournot, in fact, was quite explicit about the reality and importance of epistemic probability and advocated a rather clearly dualistic view (see Chap. 3).

In the nineteenth century the frequency theory is associated most prominently, after Venn, with C. S. Peirce, and in the twentieth century with the positivist philosophers Richard von Mises and Hans Reichenbach. Karl Popper, who valiantly attempted to repair the defects of Mises' theory, will be considered in this section, though he ultimately abandoned the frequency theory in favor of the propensity interpretation of probability.

6.2.1 Venn

Venn, whose *Logic of Chance* was first published in 1866, was the first to embark on a whole program of interpreting probability in frequency terms, expressly to avoid the pitfalls of classical statistical inference. Yet his writing still very much indicates his time. For Venn began by deplored the near total neglect of the philosophical aspects of probability in comparison with the mathematical. He went on to say quite clearly:

The opinion that Probability, instead of being a branch of the general science of evidence which happens to make use of mathematics, *is* a portion of mathematics, erroneous as it is, has yet been very disadvantageous to the science in several ways. (1888, p. vii)

He was obviously pleased with himself for carefully avoiding all mathematical symbols and formulas; even the title of his book, in contrast with innumerable authors of a *Calcul des Probabilités*, suggests more traditional than revolutionary concerns.

Like a good latter-day positivist, however, Venn rejected Bernoulli's Theorem as "one of the last remaining relics of Realism" (p. 92), on the grounds that there is no such thing as an "objective" (i.e., prior) probability. The only proper basis of a probability, he said, is a series, some of whose members possess a given attribute in a proportion that approaches a fixed limit as the series is prolonged without end. He admitted some principle of indifference, but wanted to justify it by appeal to a rough correspondence with the data of experience.

In many cases it is undoubtedly true that we do not resort to direct experience at all. If I want to know what is my chance of holding ten trumps in a game of whist, I do not enquire how often such a thing has occurred before. If all the inhabitants of the globe were to divide themselves up into whist parties they would have to keep on at it for a great many years, if they wanted to settle the question satisfactorily that way. (p. 75)

Throughout his exposition, Venn, like Ellis, was at pains to emphasize the hazards of application in science, of falsely analogizing natural series as the products of games of chance. This is one of the reasons for his rejection of the "objective" probabilities assumed by Bernoulli's Theorem. Yet in recognizing the changeableness and unpredictability of many natural series even in the aggregate, he confronted the fact that his definition of probability was strictly applicable to practically nothing except games of chance—a limitation which he disparaged in other writers on the subject. Taking the bull by the horns, he declared that what we should do with an intractably irregular natural series is simply to change it, to make it work.

For how can we have a "limit" in the case of those series which ultimately exhibit irregular fluctuations? When we say, for instance, that it is an even chance that a given person recovers from the cholera, the meaning of this assertion is that in the long run one half of the persons attacked by that disease do recover. But if we examined a sufficiently extensive range of statistics, we might find that the manners and customs of society had produced such a change in the type of the disease or its treatment, that we were no nearer approaching towards a fixed limit than we were at first. The conception of an ultimate limit in the ratio between the numbers of the two classes in the series necessarily involves an absolute fixity

of the type. When therefore nature does not present us with this absolute fixity, as she seldom or never does except in games of chance (and not demonstrably there), our only recourse is to introduce such a series, in other words, as has so often been said, to substitute a series of the right kind. (1888, pp. 164–165)

We see in the cholera example one of the distinguishing marks of a frequency theorist: the assertion that the *meaning* of the probability of an isolated event is its relative frequency of occurrence, when embedded in some chosen series. Venn did acknowledge that probabilities, in the sense of relative frequencies, are correlated with degrees of belief; but the latter, for him, not being susceptible of numerical measurement and being psychological rather than logical or mathematical, lay beyond the province of probability. He failed to notice, however, what problems he ran into when he abandoned all notions of degrees of belief and cast his theory entirely in terms of relative frequencies. For he abandoned therewith his means of making probability a philosophical, instead of merely a mathematical, concern. And indeed he did sometimes speak as though considerations of probability gave us ground for expecting one alternative in preference to another.

Whatever the limitations of Venn's technical achievements, his primary historical function was enormously to have assisted in the discrediting of the classical theory of statistical inference, which permitted “inverse probability” statements about the value of an unknown binomial parameter p . His influence was most directly noticed in Fisher's work, although, ironically, Venn's criticism of inverse probability was so grossly overstated that even Fisher felt obliged to come to its defense:

It seems that . . . Venn was to such an extent carried away by his confidence that the rule of induction he was criticizing was indefensible in many of its seeming applications, and by his eagerness to dispose of it finally, that he became uncritical of the quality of the arguments he used. (Fisher, 1956/1973, p. 28)

On the positive side, Venn's contribution amounted to at least a vision of a thoroughgoing frequency theory that would redeem the scientific value of probability. His somewhat paradoxical program—a logical or philosophical presentation of the frequency theory of probability—was picked up again by C. S. Peirce.

6.2.2 Peirce

Peirce published no books in his lifetime, but a few years after Venn's book first appeared, he began writing articles on probability and induction for *Popular Science Monthly*, while he was in the Coast Guard. He began, as did Venn, by placing probability in the domain of logic, rather than mathematics; and his introduction, again like Venn's, would scarcely suggest that a frequency interpretation was going to be offered. “The theory of probabilities,” he said, “is simply the science of logic quantitatively treated” (Peirce, 1878a, p. 606).

Peirce is important primarily for his discussion of the issue of singleton probabilities, which is one of the most lucid and searching up to his time, or since.

The idea of probability essentially belongs to a kind of inference which is repeated indefinitely. An individual inference must be either true or false, and can show no effect of probability; and, therefore, in reference to a single case considered in itself, probability can have no meaning. Yet if a man had to choose between drawing a card from a pack containing twenty-five red cards and a black one, or from a pack containing twenty-five black cards and a red one, and if the drawing of a red card were destined to transport him to eternal felicity, and that of a black one to consign him to everlasting woe, it would be folly to deny that he ought to prefer the pack containing the larger proportion of red cards, although, from the nature of the risk, it could not be repeated. It is not easy to reconcile this with our analysis of the conception of chance. But suppose he should choose the red pack, and should draw the wrong card, what consolation would he have? He might say that he had acted in accordance with reason, but that would only show that his reason was absolutely worthless. And if he should choose the right card, how could he regard it as anything but a happy accident? (1878a, pp. 608–609)

The frequency definition of probability can justify a decision only by recourse to the long run; and the problem with the long run is that, for any given individual, it is finite. In a lifetime all our inferences based on the long-run justification might turn out to be the wrong ones. The very idea of probability and of reasoning, Peirce tells us, rests on the assumption that the number of inferences made is indefinitely great. He did not shrink from the implications:

It seems to me that we are driven to this, that logicality inexorably requires that our interests shall *not* be limited. They must not stop at our own fate, but must embrace the whole community. This community, again, must not be limited, but must extend to all races of being with whom we can come into immediate or mediate intellectual relation. It must reach, however vaguely, beyond this geological epoch, beyond all bounds. He who would not sacrifice his own soul to save the whole world, is, as it seems to me, illogical in all his inferences, collectively. Logic is rooted in the social principle. (pp. 610–611)

Now, it is not necessary for logicality that a man should himself be capable of the heroism of self-sacrifice. It is sufficient that he should recognize the possibility of it, should perceive that only that man's inferences who has it are really logical, and should consequently regard his own as being only so far valid as they would be accepted by the hero. So far as he thus refers his inferences to that standard, he becomes identified with such a mind.

This makes logicality attainable enough. Sometimes we can personally attain to heroism. The soldier who runs to scale a wall knows that he will probably be shot, but that is not all he cares for. He also knows that if all the regiment, with whom in feeling he identifies himself, rush forward at once, the fort will be taken. In other cases we can only imitate the virtue. The man whom we have supposed as having to draw from the two packs, who if he is not a logician will draw from the red pack from mere habit, will see, if he is logician enough, that he cannot be logical so long as he is concerned only with his own fate, but that that man who should care equally for what was to happen in all possible cases of the sort could act logically, and would draw from the pack with the most red cards, and thus, though incapable himself of such sublimity, our logician would imitate the effect of that man's courage in order to share his logicality. (p. 611)

Now, logic is ordinarily regarded as a rather primitive discipline; it is extraordinary to make an appeal beyond logic itself for something to ground it. Kneale (1949) agrees with Peirce that “The general policy of acting on considerations of

probability in all cases of a certain kind cannot be justified by a claim that it inevitably leads to success in any finite run, however long" (p. 166). But he goes on to protest: "It is surely false, however, that the possibility of rational action in the circumstances we are considering depends on the prospects of survival of the human race" (p. 166). Peirce himself embraced the implications without qualm; he even found traditional grounds for his ultimate appeal:

It may seem strange that I should put forward three sentiments, namely, interest in an indefinite community, recognition of the possibility of this interest being made supreme, and hope in the unlimited continuance of intellectual activity, as indispensable requirements of logic. Yet, when we consider that logic depends on a mere struggle to escape doubt, which as it terminates in action, must begin in emotion, and that, furthermore, the only cause of our planting ourselves on reason is that other methods of escaping doubt fail on account of the social impulse, why should we wonder to find social sentiment presupposed in reasoning? As for the other two sentiments which I find necessary, they are so only as supports and accessories of that. It interests me to notice that these three sentiments seem to be pretty much the same as that famous trio of Charity, Faith, and Hope, which, in the estimation of St. Paul, are the finest and greatest of spiritual gifts. Neither Old nor New Testament is a textbook of the logic of science, but the latter is certainly the highest existing authority in regard to the dispositions of heart which a man ought to have. (p. 612)

Few frequency theorists could be said to have pushed any further the implications of a frequency interpretation of the probability of the single case.

Peirce was also clear that his frequency definition excluded probabilistic statements of conclusions, hypotheses, or laws of nature. Any such statement, in his view, implied a multitude of prior arrangements or universes in which the hypothesis or law might or might not be true, and he several times rejected such a notion. In his most famous passage bearing on the assignment of probabilities to natural arrangements or given universes, he said:

The relative probability of this or that arrangement of Nature is something which we should have a right to talk about if universes were as plenty as blackberries, if we could put a quantity of them in a bag, shake them well up, draw out a sample, and examine them to see what proportion of them had one arrangement and what proportion another. But, even in that case, a higher universe would contain us, in regard to whose arrangements the conception of probability could have no applicability. (1878b, p. 714)

Peirce was neither an obscure nor an uninfluential figure in philosophy, so it is interesting that his contributions to probability theory were so largely neglected. His rejection of the idea of attaching probabilities to hypotheses, with its discouraging implications for the concept of statistical inference, offers itself as a ready explanation.

6.2.3 *Richard von Mises*

I have already suggested that, by the beginning of the twentieth century, the way was prepared philosophically for a thoroughgoing frequency theory of probability. Heidelberger (1987) and Kamlah (1987) suggest that recent scientific developments

also scaffolded the construction of a formal frequency theory. Maxwell had drawn the inspiration for his statistical thermodynamics from Quetelet; his theory, plus the concepts of Brownian motion, radioactive decay—even the belated rediscovery of Mendel’s work in genetics—all contributed to the idea that probability refers to random mass phenomena.² In addition, Heidelberger believes that the crisis in mechanics attending the quantum theory supported a shift from an epistemic to an ontic conception of probability. Fechner had theoretically laid the groundwork for Mises with his metaphysical indeterminism, but the concept of probability at that time was still too closely tied to knowledge—as is clear in the theories of Venn and Peirce—for Fechner’s work to be assimilated to it.

Even as science and philosophy converged with respect to generating conditions for a frequency theory, the makings of a certain paradoxical tension also become visible. The principal thrust of positivism was the escape from metaphysics, yet the avoidance of the epistemic or psychological aspects of probability and the search for an ultimate grounding in empirical data inevitably had a metaphysical pull to it, as will become evident in the pages that follow.

As a major exponent of logical positivism, Richard von Mises advocated a strict frequency interpretation as a means of extirpating subjectivism and psychologism from probability theory root and branch. The specific goals and concerns of his theory were thus (a) to avoid reference to the principle of indifference in defining or assessing probabilities, (b) to specify the kind of series with reference to which probability is to be defined, (c) to show that such a definition has relevance for practical application, and (d) to consider the relation of this technical definition of probability to the everyday meaning, especially with reference to single-case probability.

The first of these objectives would presumably be achieved with the second. Mises defined probability by reference to an ideally constructed sequence, which he called a “collective,” defined in turn by the properties of randomness and convergence. The condition of randomness, as Mises formulated it, meant that a sequence, in order to qualify as a collective, must exclude all gambling systems. In other words, it must always be possible to defeat any prediction scheme by extending the series long enough. The condition of convergence specified that the relative frequency of a given element in the string (e.g., heads) must approach a single definite limit.

Popper (1934/1968) soon pointed out, unfortunately, that Mises’ collectives can be proved not to exist, because their two defining characteristics—randomness and convergence to a limit—are contradictory. Mathematical theorems about limits apply only to series constructed according to a rule, but any such rule could be used as a “gambling system,” and Mises defined randomness by denying the efficacy of any gambling system.

Setting aside that detail, we may consider whether Mises succeeded in avoiding use of the principle of indifference. His formulation of Bayes’ Theorem is

² It is also possible, of course, that the rediscovery of Mendel’s work around 1900 was prompted in part just by the sudden pervasive interest in random mass phenomena at that time.

instructive. Bayes' Theorem is, of course, a perfectly valid theorem under any interpretation of probability; Mises had simply to give a frequency interpretation to the prior probabilities in the formula. To that end, he asked us to imagine an urn containing nine different kinds of "stones" with a 6 on one or more sides. Some of the stones give a 6 with probability 0.1; for the other eight kinds, the probabilities of a 6 are 0.2, 0.3, ..., 0.9. Then, Mises said: "If there is an equal number of each kind in the urn, we can assume that the probability of drawing a stone from any of the nine categories is the same" (1928/1957, p. 119). This assumption gives us our initial distribution of p for a crude, discrete case, analogous to Bayes' set-up of tossing a ball on a square table. It is hard to see, however, how this assumption avoids the use of the principle of indifference, which Mises disparaged in other writers. Bayes, as he has usually been interpreted, argued by analogy from a uniform distribution of probability for the location of a ball on a table to a uniform distribution of probability for an unknown probability (see Chap. 4); Mises argued to the same result from equal numbers of stones of different kinds in a bag—while explicitly rejecting the Laplacean definition of probability in terms of a priori considerations of symmetry.

He proceeded to let himself off the hook in practical consequences by letting the number of rolls of the dice, or stones, become large. It is a feature of Bayes' Theorem that when the number of data becomes large, their weight, expressed in the likelihood, swamps the prior probability in its effect on the posterior estimate (see Chap. 8); hence with a large number of observations, variations in assumed prior distributions make relatively little difference in our final estimate. This is the feature which led him to refer to Bayes' Theorem as the Second Law of Large Numbers. It apparently allowed him to claim—though he did not do so explicitly—that his reliance on the principle of indifference, or the Laplacean definition of probability, was unimportant practically, even if it were technically offensive.

With regard to application, Mises confronted the same problem as Venn. Obviously there are some series which are not suitable for defining probability. Bernoulli's Theorem allows the possibility, for example, that, in an indefinitely long series of tosses, we should get all heads, or nearly all heads; it merely assures us that there is a very small probability of a nonconvergent series. But whereas Bernoulli's Theorem establishes convergence *in probability*, Mises needed convergence for sure. Hence a collective is not just any sequence that might be generated by tossing a coin, but one of a special set of series. It could be said that in postulating the existence of a limit in the collective, Mises was in effect assuming Bernoulli's Theorem as true a priori and using it in his definition of probability—thus achieving by sheer stipulation what took Bernoulli 20 years to prove.

It is not quite true, however, that Bernoulli's Theorem was simply assumed, for that theorem began with classical probabilities, defined by the principle of indifference. It is just those a priori probabilities which Mises rejected; and in making relative frequency the *definition* of probability, he made Bernoulli's Theorem circular at the same stroke: It now merely said that the relative frequency of occurrence of an event in a series approaches its relative frequency of occurrence. From the frequentist perspective, of course, the trouble lies with the a priori, "metaphysical" character of the classical probabilities: If it is only through an infinite set of experiments

that it could be established that the probability of a head with a coin was $\frac{1}{2}$, then Bernoulli's Theorem is necessarily circular.

In general, the problem in using the frequency theory is that we never have the necessary data for the measurement of probabilities. As Venn (1888) observed, we do not—and cannot—ascertain numerical probabilities for games of chance by resorting to infinite sequences of trials. In order to assert that the probability of a head with a given coin is $\frac{1}{2}$, we must simply “substitute a series of the right kind”; that the limit of the series that would be generated by this coin would be $\frac{1}{2}$ becomes entirely a matter of faith. The frequency theory thus ends up with an inversion of its professed aim: Setting out to put probability on a purely empirical basis, it defines probability with respect to ideally constructed sequences; and these would appear to be as difficult to connect with empirical statements as the classical a priori probabilities.

The infinity of trials itself necessary to verify any probability statement also thwarts the empirical ambitions of the frequency theorists. They have typically been committed to the verifiability theory of meaning, which would hold all frequency probability statements to be meaningless. The recourse can only be to approximations, and this is the route Mises took, but it leads us in a circle. If we ask how close the approximation must be to be acceptable, we are involved essentially in a significance test. We can only give a probabilistic answer, so that we are defining probability in terms of probability.

The problem of applying the frequency theory, however, goes beyond the question of how adequately finite approximations can model infinite series. Consider one of Mises' own examples, of there being an 80% probability of a certain tennis player winning a particular tournament in London. He said, even in this case, that the 80% probability referred to a collective, but he did not say what the elements of this collective were. “In order to apply the theory of probability we must have a practically unlimited sequence of uniform observations” (p. 11); but it is hard to guess what sequence he had in mind: a thousand (or a dozen?) previous tournaments by this player? by other players in his class? tournaments in London? against a particular opponent? or class of opponents? Quite possibly a particular sequence could be defined by a suitable refining of the question; but the question as it stands—“What is the probability of this tennis player winning this tournament?”—is not well defined in Mises' terms; and once we selected a reference sequence, we would still face the problem of determining—or assuming—that the sequence was random and that the ratio of wins approached a limit. In many other examples where we might refer to probability in ordinary speech, it will be even more difficult to define an appropriate collective. And, of course, Mises, like Peirce, rejected, as unscientific, statements of the probability of hypotheses or laws, or of any singular event. “It is utter nonsense to say, for instance, that Mr. X, now aged forty, has the probability 0.011 of dying in the course of the next year” (pp. 17–18).

Mises recognized clearly that his definition of probability excluded many of the instances of ordinary language, but he offered a robust defense of the restricted interpretation, comparing it with the scientific concept of work. The physical concept of work, as the product of force and distance, is obviously inapplicable to the

work done by an actor in reciting his part in a play, or by a mathematician in solving a differential equation, and we would not want to deny that either of these people might be working very hard. The solution to which we are all long accustomed in this case is simply to demarcate a particular sphere of application for the scientific concept, and to recognize that the original, everyday concept of work, while still related to the physical concept, embraces a much less restricted meaning. Mises argued that exactly the same situation should obtain with respect to probability, that attempts by some theorists to include, in a scientific concept of probability, statements, for example, about the probability of winning a battle (or a tournament?), are simply misguided, that we should not even try to bring all popular instances of probability statements into a scientific conception. In the last analysis, then, the concept of probability is no more problematic, Mises contended, than the concept of work.

This last argument is probably the strongest of his positions; it is at least a legitimate, meaningful position to take. In other respects—the coherence of his definition of probability, of his ideas about application, of his rejection of the principle of indifference, and of his advocacy of empiricism—Mises' achievements were perhaps less impressive than his influence.

6.2.4 Reichenbach

Hans Reichenbach, *Theory of Probability* (1935/1949), was motivated at least in part by the aim of repairing defects in Mises' theory. He was especially concerned that his theory comprehend *all* instances of *probability* in a frequentist sense. Thus whereas Mises' definition led him to exclude many ordinary usages of *probability* from his theory, Reichenbach claimed to his advantage that his theory “leads to a meaning of the term that makes the usage of language conform to human behavior” (1935/1949, p. viii). In introducing his theory, he wrote that “a final theory of probability that satisfies both mathematical and logical requirements can now be presented” (p. v). Despite the fact that his formulation of the frequency theory is the most extreme, it was also his intention, in contrast to Mises, to keep the epistemic reference central, with its concerns for inductive inference. His proposed range of applications for Bayes' Theorem was “extremely wide” (p. 94), nearly as ambitious as Price's: medical diagnoses, historical explanations, detective work, and so on. He also claimed to prove that all inductive inference is reducible to induction by enumeration.

Although Reichenbach's work came too late to exert any formative influence on either Fisher or Neyman and Pearson in their theories of statistical inference (Chap. 7), it is still worth examining to two reasons: (a) Reichenbach's theory comes closest to duplicating the everyday, unreflective meaning of probability (cf. Chap. 10), and (b) we might get a sense of how the Fisher and Neyman-Pearson theories could be improved.

The technical aspects of Reichenbach's definition of probability are not markedly different from Mises'. The sequences taken as *definientia* for probability are

ideally constructed series, which Reichenbach called “normal sequences”; they represent his attempt to weaken Mises’ condition of randomness. Adopting the causal terminology of freedom from aftereffect in constructed sequences, he defined a sequence as normal “if it is free from aftereffect and if the regular divisions belong to its domain of invariance” (p. 144; original in italics). Freedom from aftereffect means that the relative frequency composition of a series is unaffected by selecting out all the predecessors (or successors) of a given element (e.g., the zeros), a given pair (e.g., 01), and so on. The latter condition means that the series must be unchanged in the relative frequencies of its elements by a selection of elements at regular intervals of any given length.

His definition represented a certain advance over Mises’, in that, where Mises’ collectives could be proved not to exist, Reichenbach’s normal sequences merely could not be proved to exist (Popper, 1934/1968). His escape from the principle of indifference was not much more successful, however. Not that his rejection of it lacked vigor: “To transform the absence of a reason into a positive reason represents a feat of oratorical art that is worthy of an attorney for the defense but is not permissible in the court of logic” (p. 354). He admitted that the principle of indifference possessed plausibility in many of the situations in which it is traditionally applied, but he set about to explain its plausibility by substituting positive arguments, which were basically the familiar appeals to considerations of symmetry. He gave a geometrical example involving spins of a roulette wheel, where the ball is assumed to come to rest at some point along the continuous circumference of the wheel. He disparaged the Laplacean assumption that equal areas correspond to equal probabilities, but he still used that assumption in his own solution; it was merely hidden under an elaborate discussion of constraints on the behavior of the probability distribution function in the transition from discrete to continuous probability (pp. 355–359). He rejected the principle of indifference as committing the “fallacy of incomplete schematization” (essentially a failure to specify all relevant conditions); but in his own attempt at supplying positive, *a posteriori* arguments for equi-probability assumptions, he did not address himself to the problems described by Mises in the case of equiprobable outcomes at dice.

The logical character of this theory comes out in his attempt to understand the probability of single events. Essentially, his strategy was to embed the singular event in a reference class and then to identify the probability with the relative frequency in that class. More precisely,

There exists only one legitimate concept of probability, which refers to classes, and the pseudoconcept of a probability of a single case must be replaced by a substitute constructed in terms of class probabilities.

The substitute is constructed by regarding the individual case as the limit of classes becoming gradually narrower and narrower. . . . A repeated division of the main sequence into subsequences will lead to progressively better results as long as the probability is increased at each step. According to general experience, the probability will approach a limit when the single case is enclosed in narrower and narrower classes, to the effect that, from a certain step on, further narrowing will no longer result in noticeable improvement. (pp. 375–376)

What Reichenbach presumably intended to say is, not that the probability is increased, but that it becomes more stable, as the reference class becomes narrower. As any real reference class grows more specific, however, the number of its members decreases, and so does the denominator of the fraction representing the relative frequency of occurrence of the event in question. But then the sequence of fractions, instead of approaching 1, 0, or any other value in a steady fashion, jumps around more and more erratically (e.g., 4/6, 3/5, 2/4, 2/3, 1/2, 1/1), landing eventually on 1 in the limiting case of the individual (cf. Lucas, 1970, pp. 104–107).

Reichenbach recognized that selection of an appropriate or best reference class posed a problem for his theory, as for any frequency conception. He first recommended selection of the narrowest class for which reliable statistics could be compiled, but he acknowledged that this rule, in turn, left some ambiguity. In any event, he ended up with the conclusion that the probability of a single case depends upon the state of our knowledge, a consequence which does not obtain for probabilities referred to classes.

To speak of the *meaning* of the probability of a single event being a relative frequency is still a bit of a strain; that is why Reichenbach called it a “substitute” meaning above.

I regard the statement about the probability of the single case, not as having a meaning of its own, but as representing an elliptic mode of speech. In order to acquire meaning, the statement must be translated into a statement about a frequency in a sequence of repeated occurrences. The statement concerning the probability of the single case thus is given a *fictitious meaning*, constructed by a *transfer of meaning from the general to the particular case*. The adoption of the fictitious meaning is justifiable, not for cognitive reasons, but because it serves the purpose of action to deal with such statements as meaningful. (pp. 376–377)

Probability statements about individual cases cannot, strictly speaking, be asserted, as subject to judgments of truth or falsity; instead they are *posited*, a posit being understood as something like a wager: “A posit is a statement with which we deal as true, although the truth value is unknown” (p. 373; original in italics).

If the “truth value” of a posit is known, Reichenbach called it an “appraised posit,” and this concept led him to suggest an alternative interpretation of the probability of a single case, which he actually preferred. The idea was originally put forth by Boole (1854), who observed that probability statements could be taken to refer either to events or to statements about events. Reichenbach’s logical interpretation of probability identified a probability with the truth value of a posit, which he called its weight. Generalizing this logical interpretation of probability, he proceeded to construct a probability logic. In his system, the objects of probability were propositional sequences, which were Whitehead’s propositional functions with ordered sequences of events as arguments (the ordering is irrelevant for finite sequences); and the probability of a single proposition was given meaning in a way analogous to that in the frequency interpretation: “The numerical value of the frequency in the sequence is transferred to the individual statement in the sense of a rating, although the individual statement taken alone exhibits no features that could be measured by the rating” (p. 381). In an infinite propositional sequence, the

probability, as the relative “truth-frequency” of the sequence, can take on a continuous range of values from 0 to 1; but a series of length n yields an $(n + 1)$ -valued logic. “Truth and falsehood can thus be regarded as the limiting cases of probability resulting when the sequence is reduced to one element” (p. 398).

The concept of probability as the limiting truth-frequency of a propositional sequence suffers, however, from some critical ambiguities. The first problem, as ever, is how the reference sequence is to be chosen, but there are other problems as well. Suppose $3/4$ of the hypotheses in a series are known to be true and the remainder are known to be false. Then it appears that we should have to assign all of them, as members of a particular propositional sequence, a probability of $3/4$, rather than 1 or 0 individually. Finally, suppose we knew the truth status of all the hypotheses or propositions in a long sequence. According to Reichenbach, we have to know this much to assign a probability to a hypothesis; but if we knew the truth or falsity of all these propositions, it is not clear why we would still need to know this one probability.

We should note in particular how Reichenbach proposed to handle the matter of judging the probability of hypotheses or theories. He saw no problem in principle:

Should we some day reach a stage in which we have as many statistics on theories as we have today on cases of disease and subsequent death, we could select a reference class that satisfies the condition of homogeneity . . . , and the choice of the reference class for the probability of theories would seem as natural as that of the reference class for the probability of death. In some domains we have actually been witnesses of such a development. For instance, we know how to find a reference class for the probability of good weather tomorrow, but before the evolution of a scientific meteorology this reference class seemed as ambiguous as that of a scientific theory may seem today. The selection of a suitable reference class is always a problem of advanced knowledge. (p. 440)

If, on the other hand, we want to judge or compare the probabilities of historical statements, for instance, that Caesar stayed in Britain, that Nero ordered the burning of Rome, or that Bacon wrote Shakespeare’s plays, “We investigate the number of chroniclers who report such a fact; and we measure the reliability of the individual chronicler by the number of his reports that are corroborated by reports of other writers” (p. 379). Reichenbach is startlingly close here to some of the naive work on probability of testimony of several witnesses, done in the early years of the development of the mathematical theory.

Cournot (1843) had already set forth very clearly the objections to these efforts a century earlier. Like Poisson, he took a special interest in jurisprudence and the statistics of civil and criminal cases. Cournot, who was generally surpassed only by Keynes in his circumspection, argued, essentially—to use Pepper’s (1942) terms—that structural corroboration counts more than multiplicative:

We believe firmly in the existence of the person named Augustus, not only because a whole crowd of historians have spoken of him and are agreed on the principal circumstances of his life, but also because Augustus is not an isolated personage, and the record of his life makes sense of a host of contemporaneous and subsequent events which would lack any basis and would no longer have any connection, if so important a link in the historical chain were removed.

If some particular individuals were to doubt the Pythagorean theorem or the existence of Augustus, our belief in them would not be shaken in the slightest: we should not hesitate to conclude that they were disordered in some of their intellectual faculties, that they had gone quite beyond the normal conditions in which our faculties must function, to fulfill their proper purpose.

It is thus not solely on repetition of the same judgments, nor on unanimous or nearly unanimous agreement, that our belief in certain truths is based: it rests principally on the perception of a rational order linking these truths, and on the conviction that the causes of error are anomalous, irregular, subjective causes, which could not give birth to such a regular and objective coordination. (1843, pp. 421–422)

Cournot concluded with a caution which evidently escaped Reichenbach's notice (there is, in fact, no reference to him in the index to Reichenbach's *Theory of Probability*):

We shall resist the temptation to apply the calculus to the probability of facts supposedly known through a chain of witnesses, or through tradition. Not only are values in no way assignable to the elements entering into such calculations; but the very combinations of these elements in the calculus rest on gratuitous hypotheses, by which is established a fictitious independence between events that are really all of a piece, and whose mutual dependence confounds any legitimate application of the theory of chances. (p. 415)

When Reichenbach went on, finally, to say that in many kinds of everyday situations, probabilities may be estimated by a “rational reconstruction” of betting behavior, it becomes plain that he had not entirely succeeded in banishing psychologism and subjectivism from probability theory.

Reichenbach's theory is bizarre in being at once so contrived and so naive. It is hard to say that his theory has been very influential, in the ordinary direct sense, but it is important in representing in a global way the unarticulated popular notions on the subject: The everyday meaning carries the strong suggestion that all probabilities are interpretable in both frequency and epistemic terms, and Reichenbach's is the only theory to agree. Indeed, one precocious 11-year-old I interviewed, struggling for the first time with the perplexities of single-case probability, came up with language startlingly close to Reichenbach's (Chap. 10).

6.2.5 Popper

Yet another century after Comte, Popper (1934/1968) was still trying to rid science of metaphysics; his criterion of falsifiability as the demarcation between science and metaphysics is probably his best-known contribution to the philosophy of science. Since it also left him, with the positivists, still in the arms of the analytic-synthetic dichotomy, he naturally opted for an “objective” (i.e., frequency) theory because he believed it to be the only one which would make the probability calculus applicable to empirical science. His frequency theory was put forth in Chap. 7 of

The Logic of Scientific Discovery (the German edition of which was published in 1934) and in two long strings of appendices and footnotes.³

Having criticized both Mises' and Reichenbach's technical definitions of probability, Popper was obliged to try one of his own. Mises' criterion of randomness he replaced with the requirement of " n -freedom." As in Reichenbach's normal sequences, the relative frequency in an n -free series must be insensitive to selection according to arbitrary n -tuples. Popper gave a method (in footnotes *1 and *2 of appendix iv) for constructing such n -free sequences, thus proving that such series existed. Mises' axiom of convergence he first replaced with his "requirement of uniqueness"—that there be one and only one probability (i.e., relative frequency) associated with a given n -free binary sequence. Later (footnote *2 to §64) he replaced the requirement of uniqueness with the "requirement of finitude: the sequence must become, from its commencement, as quickly n -free as possible, and for the largest possible n " (p. 187). The latter requirement was loose enough not to contradict the criterion of n -freedom, but Popper was of course left with the same question as all the other frequentists: How can probability theory apply to any real phenomenon, if it pertains only to a very special sort of ideally constructed sequence? Clearly the empirical sequences to which we might want to apply the concept of probability are not constrained to follow the rules of anybody's particular theory. Ultimately, as was noted above, he resolved the problem by abandoning the frequency theory for the propensity interpretation of probability.

Like Venn, Peirce, and Reichenbach, Popper was interested in probability for its role in epistemology, not just in mathematics. Like Mises, however, he reserved the scientific concept for frequency applications. To refer to the so-called probability of hypotheses, he used the terms "logical probability" and "corroboration," which he took care to distinguish from each other. "Corroboration" he used in a special sense connected with falsifiability: A hypothesis is said to be corroborable to the degree to which it is subject to stringent tests, and hence, by implication, the corroboration of a hypothesis is proportional to the extravagance of its claims. Tautologies are the lowest of statements in corroboration, since they cannot be falsified at all, but they are at the same time highest in (logical) probability. The two concepts are complementary, and Popper maintained that it is corroboration in which scientists are mainly interested. If high probabilities were our goal, he suggested, we should endeavor to say as little as possible; we should always choose the most ad hoc hypotheses. Instead, he said with a perfectly straight face, when faced

³ Some impatient readers have suggested that perhaps these appendices could be surgically removed, but their deletion would unquestionably be deleterious to any appreciation of Popper's work. Whatever the merits of its specific contents, *The Logic of Scientific Discovery* is a document of rare value in the history of ideas. There are few thinkers around who have been willing to leave their youthful ideas intact for all posterity to see, as they merely appended their changes of mind with asterisks. Most have been content to frustrate ontogenetic comparisons by letting their previous mistakes disappear quietly from print, to be replaced with pat, new, authoritative versions. Taken all together the book by its 1968 edition exhibits a predictable progression from more to less brash claims for his theory, and for the humility which made possible his greater gift to the science of intellectual development, Popper is to be commended.

with competing theories, we should “hold on, for the time being, to the most improbable of the surviving theories” (1934/1968, p. 419; original in italics), i.e., the one that can be most severely tested.

He still faced the question of the relation between corroborability, logical probability, and relative frequencies, having just moved it a step further back. Popper believed that it was possible to develop a metric for logical probability (i.e., the probability of hypotheses, as distinguished from the probability of events) and that relative frequencies, being the more fundamental concept, could supply the basis for measurement. (He argued that a metric for logical probability could never be based on sheerly logical considerations, for then it would be subject to modification in individual cases by empirical data, contradicting its logical status.) He went so far (in appendix *ix) as to make numerical calculations of logical probabilities, taking as a basis a “Laplacean distribution,” i.e., the same principle as Laplace used in deriving the Rule of Succession. When later he decided to accept the idea of a logical metric for “secondary” probabilities, which are probabilistic statements about probability assertions, he began to wonder himself what had happened to the frequency theory. In a new addition to appendix *ix (pp. 418–419), he ultimately acquiesced to the identification of degree of corroboration with logical probability, provided, he insisted, that the assessment was based on our sincerest effort to overthrow the hypothesis. Thus he tortuously argued his way to statements about hypotheses, based on relative frequencies, or at least on classical probabilities; the force of his original intent was relegated to the side constraint that we have made every effort to falsify the hypothesis; and he ended up somewhat closer to Reichenbach.

6.3 The Concept of Randomness

Popper’s series, like Mises’ collectives and Reichenbach’s normal sequences, provide an outcome, or *achievement*, criterion of randomness, to use Werner’s (1937) distinction; yet in applying the theory, we generally use probability to model random *processes*. The two criteria, or conceptions, of randomness are not the same; in fact, they are to some extent opposed. There are thus problems in the application of the frequency theory, having to do with ambiguity in the concept of randomness, which appear to be untouched by Mises’ remarks about every scientific model being an approximation.

The English word *random* derives from the French *randir*, to run fast or gallop, and originally connoted impetuosity or lack of control.⁴ The sense of unpredictability, of freedom from constraint or governance by a rule, fits the conception of randomness as characteristic of a process, or of a device or set-up which generates data strings. In many situations, however, the only way we have of judging

⁴It is interesting to note that in British dialect the word *randy* means unruly; the *Oxford English Dictionary* maintains that its origins are unclear, but Partridge (1966) draws the obvious connection.

randomness in this sense is by inspecting the data strings (“reading the signs”) rather than the device itself; this circumstance leads us to the achievement conception of randomness as the absence of discernible pattern (n -freedom, etc.). We may follow Spencer Brown (1957) in referring to these two meanings as primary and secondary randomness, respectively.

Now, taking a simple example, suppose we are testing a coin for bias by inspecting a string of 100 tosses. There are many kinds of possible bias (e.g., bias in doubles or triples), but most often we are concerned with elementary bias—a preponderance of heads or tails. If the model is correct, and the coin is free of bias, then we should expect about half the tosses to turn up heads. The ratio of 1:1 strictly applies, however, only to the indefinite long run; and, if the series is to be genuinely random (in the primary sense), it must allow anything to happen on the way there. Our expectation of a roughly even split in 100 tosses is based on the assumption that this particular sample of 100 tosses is representative of the whole string. But if *any* such sample were representative, and unbiased, we should have grounds for rejecting the hypothesis of (primary) randomness; we would have evidence that something was intervening to keep the process too evenly balanced. The excessive regularity would be a predictable feature of the series, contradicting the assumption of unpredictability.

Hence we can already see the outlines of a paradox: In order for a series to be primarily random, it must (sometimes) fail to be secondarily random. If the long-run relative frequency is $\frac{1}{2}$, then strings of any finite length must sooner or later take on all other possible values. In fact, of course, we do not assume that a string of 100 heads is an exceptional output from an unbiased coin, or that it is a “chance” event; we decide there is some other factor controlling the behavior of the coin that we ought to know about. The upshot is that we are willing to call a series random or chance only if it possesses an intermediate degree of bias, if it exhibits neither too regular an alternation nor too great a preponderance of one element.

As a further consequence in turn, the field of possible series in the set of random strings is restricted, with the apparent result that our calculated probabilities are distorted. In excluding certain sequences from the set of possible random events, we are essentially treating them, as d’Alembert (1767) suggested, as impossible, as having probability zero, and thus destroying the strict additivity of probability. Eliminating the most radical deviants makes it easier for the remaining members of the set to look more extreme, as large probabilities cannibalize small ones and become even larger. We can imagine the process zipping back until the most perfectly random series is left, as the only acceptable outcome on the chance hypothesis.

The issue is not merely theoretical, but arises sometimes explicitly in research contexts. Suppose, to insure randomness, we use a random number table in assigning research participants to one of two groups. Then suppose it turns out that nearly all the men are in the same group. To avoid the obvious possibility of a confounding of the experimental treatment with sex differences, we discard that random number string and select another. We will keep on until we find one that yields an assignment that is good and random, in the secondary sense. But to follow this procedure

is tantamount to selecting the assignment we wanted in the first place—which is the antithesis of the “blindness” sought in random assignment.

Just this issue divided Fisher and Gosset: The latter argued frankly for a principle of optimum rather than random selection, to insure a nonsystematic design; Fisher prevailed with the point that randomness was essential for valid estimation of the error variance. His response appealed to the indefinite long run and entailed that flukes on the way to infinity are simply to be absorbed in their proper course:

If we . . . agree that an event which would occur by chance only once in seventy trials is decidedly “significant,” in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the “one chance in a million” will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us. (1935, p. 13)

But, as he also observed, any set of experiments may be regarded in turn as a single aggregate experiment, whose length is the sum of the lengths of the individual experiments. Spencer Brown puts these two premises together to form what he calls the Experimental Paradox: that “No series of experiments, however significant in themselves, can suffice for the experimental demonstration of any phenomenon” (1957, pp. 65–66). The conclusion is true enough, inasmuch as infinity, which is what our hypotheses are implicitly about, can always be counted on to dwarf into insignificance any observations of a finite run, which are all we shall ever make.

Spencer Brown (1957) believes the dilemma goes to the heart of the concept of randomness; he notices that it affects even the construction of random number tables. Maurice Kendall (Kendall & Babington Smith, 1939) made stringent tests on the output of his random number machine; when he left the machine in the hands of an assistant who was unaware of the tests that would have to be passed, 10,000 numbers were generated that Kendall threw out. Even so, he indicated that there were certain portions of the table that should be avoided when sampling fewer than 1000 digits. Fisher and Yates (1938), using for their tables the 15th to 19th digits in a 20-place table of logarithms, found that part of the table had too many 6s, and changed it (Hogben, 1957). In both cases a decision is involved, on formal or informal grounds, about whether the device is truly random, in the primary sense, or whether it has a real bias, i.e., one that would be expected to continue if the output were prolonged. The question is a subtle one; much effort in recent decades has gone into the search for the “Snow White randomizer” (“the fairest one of all”; H. Friedman, 1978). The decision in these two cases would obviously tend to result in an excessive regularity (in the sense of secondary randomness), which in the course of repeated application would tend to make nature appear more uniform than it is.

Popper’s series are the extreme example of the tension between primary and secondary randomness: Hardly any series qualify as random in his very highly restricted sense. To apply his theory, we must assume that empirical sequences of interest, which are ostensibly random in the primary sense, will behave like the secondarily random series covered by the theory. To make that assumption is actually to frame an hypothesis about empirical series. That hypothesis is testable, then,

if we are prepared to regard very small probabilities as actually zero. This solution, again, goes back to d'Alembert (1767), who suggested that if we saw letters on a table spelling out CONSTANTINOPOLITANENSIBUS—“*un mot connu*” (p. 293)—we would not entertain for a second the hypothesis that the arrangement was random; we would not say the probability was small, but that it was zero. D'Alembert was essentially denying what Spencer Brown (1957) would call the “monkey theorem,” from the stock example of monkeys at typewriters producing Shakespeare's plays.⁵ It is true that in the long run the improbable will surely happen and that perfect regularities may turn up which are actually due to chance; but Popper proposed that we make a methodological decision never to regard regularities as chance. The problem of falsifiability of probability statements is thus solved by the device of not allowing highly improbable events to happen, thus making some observations incompatible with our hypothesis. If they do happen, we modify our hypothesis to make what actually happened look more probable in retrospect. The claim that very improbable events do not happen, however, seems, unfortunately, to be about as metaphysical and a priori as the corresponding aspects of the classical theory, which the frequency definition was designed to escape. Moreover, the concept of “practical (or moral) certainty” involved is not an especially clear one to set at the base of an empirically oriented theory.

The problem of possible divergence is often dismissed as being of academic interest only since divergent series, in relation to the whole set of possible infinite series, have probability 0. But Fine (1973) argues that this solution is too glib.

Probability 0 does not necessarily mean a small or negligible set of possibilities. (p. 93)

The set of sample sequences for which the relative frequency does not converge to the probability is as large a set, in the sense of cardinality, as the set for which it does converge. (p. 96)

He goes on to consider possible justifications for our neglect of such contingencies. Recall that secondary randomness, such as Popper's n -freedom, was defined as the absence of discernible pattern; but the discernibility of a pattern clearly depends on the discerner. It is generally to our interest, in theory construction, to identify patterns wherever we can, for a more economical description, leaving randomness as the residue. Our probabilistic methods start to apply only where our theoretical specification leaves off and thus exist in a complementary relationship with nonstatistical methods. The usefulness of frequency probability thus depends on our finding a place to declare a halt to theoretical specifications, and, generally speaking, we are willing to leave as random only series of a high degree of complexity.

But now the unsettling implication is that the law of large numbers itself can be seen, not as a law of nature, but as a consequence of our selection. The stability of frequencies in long random series, Fine (1973) suggests, can be attributed to the kinds of series we are willing to allow into the class of random series. Our success

⁵ According to Dick Cavett, this experiment was actually once done, and, surely enough, none of the monkeys ever came up with any of Shakespeare's plays, although three of them did eventually produce novels by Jacqueline Susann.

in using relative frequency arguments can thus be said to result from our selectivity in applying it, rather than from a natural law. A further conclusion,

and one of significant practical importance, is that since the apparently convergent relative-frequencies in an empirical random sequence do not arise from, and are not reflective of, an underlying empirical law, we have no grounds for believing that the apparent limit is indicative for future data. Convergence was forced by preselection, on grounds once thought to conflict with convergence, of the data sequence and is not a reflection of some underlying process of convergence governing all trials. The appearance of stability is not, as it has been thought, evidence for the hypothesis that nature ensures stability. Our intuitive expectation that apparent trends must persist has even less justification than would appear to a skeptic familiar with Hume's firm attack on induction; the trends are of our own making. (Fine, 1973, p. 94)

To support our practice of using the secondary randomness of finite series as our test criterion, we appeal to the law of large numbers; but if Fine's argument is correct, the comfort we draw from it would appear to be curiously onanistic. If we treat the small probabilities of nonrepresentative samples as actually zero, then the fit assured by Bernoulli's Theorem is better than we thought, but only because we make it so. We may indeed find future series to be convergent just because we shall not submit deviant series to test. The same selection process protects the appearance of success.

What determines the kinds of series that will be subjected to modeling as a random process? Fine (1973) offers some interesting "pseudopsychological reflections":

How do we know, or why do we believe, that we cannot accurately predict the outcome of a vigorous toss of a balanced coin? I submit that this belief is primarily a product of personal frustration with a lack of success in actual attempts and secondarily an acceptance of the judgment of others that it cannot be done successfully. (Researchers into extrasensory phenomena and gamblers, though, may not share these beliefs.) Giving up on deterministic prediction of the outcomes of coin tosses, we then look for weaker but still informative predictions. Can we, for example, put narrow bounds on the exact number of heads in n tosses? Apparently our personal experiences indicate that we cannot correctly assert that "there will be between 40 and 60 heads in 100 tosses." So we proceed yet further along the path of looser predictions and talk about what will happen in a hypothetical long run. At this point we find ourselves making useless but safe predictions. Perhaps what we are attempting is not possible. We persist because the indefiniteness of our long run predictions makes them nearly irrefutable and reduces the incidence of frustration. (p. 104)

Part of the problem is the vagueness in our criterion for excluding certain outcomes. Without knowing the limit of a given series in advance, for example, we are unable to say whether, or to what degree, a series is converging.

This situation would be somewhat more tolerable if it could be asserted that once we actually are within ϵ of the limit we can recognize it, although we could not determine the required n in advance of observation. For example, while I do not know how long it will be before, if ever, I visit the moon, I can at least determine it when this event occurs. The long run guarantee of convergence is thus ineffective since we are unable, either before or after observation, to determine how long a "long" run must be. (Fine, 1973, p. 103)

We are accustomed to cutting through the vagueness with arbitrary criteria, though generally without having appreciated their source. Spencer Brown (1957) suggests that use of the conventional 5% significance level can be vindicated by

reference to the human scale of existence, of our lifetimes and our rate of observation. If our rate of observation were fast enough that we encountered strings of 100 heads about as often as we actually encounter strings of 5 heads, then we should naturally take such a string in stride. If, on the other hand, our lifetimes were so short that we could never observe more than a dozen of anything, a run of three or four heads might look quite significantly long. The historical 0.05 level happens to fit well on such a scale, but the use of values beyond 0.01, he says, is meaningless, “since if an experiment is at all difficult we are not likely to do it more than 100 times” (1957, p. 87). These traditional criteria receive some psychological support also from the work of Cowles and Davis (1982), who found that people on the average use implicit criteria of around $p = 0.1$ for doubt and $p = 0.01$ for disbelief.

We can contrive a pragmatic falsification, in short, only by demarcating intervals of time, or lengths of series, which we regard as relevant to us. With that conclusion, ironically, the objectivist aspirations of the frequency theory come up against a profound subjectivism at the heart of the theory: Not only does any process of application entail “metaphysical” commitments about what is physically possible—which amounts to saying that we will never allow anything too improbable to happen—but any meaningful criterion necessarily makes at least implicit reference to the human span of observation.

6.4 Summary

The frequency theory of probability was intended from the start to put the concept of probability on a surer, more empirical footing. Rejecting the principle of indifference as a priori and subjective, it attempted to substitute, for the purpose of probability assessment, the relative frequency of an event in an infinite series and to make such a relative frequency definitive of probability. The attempt failed, for a variety of reasons. In making relative frequency the operational definition of probability, the frequency theory surrendered reference or applicability to anything outside relative frequencies. It renounced all connection to knowledge or expectation, or even objective propensity, and it made Bernoulli’s Theorem a meaningless circle: The relative frequency of an event will very frequently approach its relative frequency. The series with reference to which probability is defined must be theoretically specified, but it is in the nature of empirical series that their behavior cannot in principle be so specified. The empiricism of the theory demands that probability statements be verifiable, but the reference to an infinite series of observations alone precludes verifiability. The attempt, moreover, to test probability statements entails recursive reference to probability. Constructing criteria for such a test also requires demarcating some boundaries for the decision; any such criterion can meaningfully be defined only in relation to the temporal scale of human existence, which leaves subjective components built in which are not clearly more objectionable than the subjectivity of the principle of indifference.

Twentieth-century developments in the frequency theory—Mises, Reichenbach, Popper—came too late to be influential in the original formulation of the modern theory of statistical inference by Fisher, though they had much to do with the climate of acceptance for the later version by Neyman and Pearson. At the time that Fisher appeared on the scene, Venn's (1866/1888) work was the most recent. Venn was influential in having discredited the classical theory of statistical inference, and the Laplacean concept of probability on which it was based. Since the epistemology of the eighteenth century had survived in fundamental ways unchanged into the twentieth, Venn's criticism pointed to the need for a new theory of statistical inference not based on the inverse probability of Bayes and Laplace.

References

- Bertrand, J. (1889). *Calcul des probabilités* [Calculus of probabilities]. Paris, France: Gauthier-Villars.
- Beth, E. W., & Piaget, J. (1961). *Épistémologie mathématique et psychologie* [Mathematical epistemology and psychology] (*Études d'Épistémologie Génétique*, Vol. 14). Paris: Presses Universitaires de France.
- Boole, G. (1854). *An investigation into the laws of thought, on which are founded the mathematical theories of logic and probabilities*. London, UK: Walton and Maberly.
- Comte, A. (1864). *Cours de philosophie positive* [Course in positive philosophy] (2nd ed.). Paris, France: Baillière. (1st ed., 1830–1842).
- Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités* [Exposition of the theory of chances and probabilities]. Paris, France: Hachette.
- Cowles, M., & Davis, C. (1982). Is the .05 level subjectively reasonable? *Canadian Journal of Behavioural Science*, 14, 248–252.
- D'Alembert, J. le R. (1767). Doutes et questions sur le calcul des probabilités [Doubts and questions on the calculus of probabilities]. In *Mélanges de littérature, d'histoire, et de philosophie* (Vol. 5, pp. 275–304). Amsterdam, The Netherlands: Chatelain.
- Daston, L. J. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.
- Falmagne, R. J. (1975). *Reasoning: Representation and process*. Hillsdale, NJ: Erlbaum.
- Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1, 2nd ed.). New York, NY: Wiley.
- Fine, T. L. (1973). *Theories of probability*. New York: Academic Press.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). New York, NY: Hafner. (1st ed., 1956).
- Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Edinburgh, Scotland: Oliver and Boyd.
- Friedman, H. (1978). The Snow White randomizer: Simple, economical, and the fairest one of all. *Bulletin of the Psychonomic Society*, 12, 227–228.
- Gottinger, H. W. (1974). Review of concepts and theories of probability. *Scientia*, 109, 83–110.
- Hacking, I. (1975). *The emergence of probability*. Cambridge, UK: Cambridge University Press.
- Heidelberger, M. (1987). Fechner's indeterminism: From freedom to laws of chance. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 117–156). Cambridge, MA: MIT Press.
- Hogben, L. (1957). *Statistical theory: The relationship of probability, credibility and error*. New York, NY: Norton.

- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York, NY: Basic Books.
- Kamlah, A. (1987). The decline of the Laplacean theory of probability: A study of Stumpf, von Kries, and Meinong. In L. Krüger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution. Vol. 1. Ideas in history* (pp. 91–116). Cambridge, MA: MIT Press.
- Kendall, M. G., & Babington Smith, B. (1939). *Tables of random sampling numbers*. Cambridge, UK: Cambridge University Press.
- Keynes, J. M. (1973). *A treatise on probability*. New York, NY: St. Martin's Press. (Original work published 1921).
- Kneale, W. (1949). *Probability and induction*. Oxford, UK: Clarendon Press.
- Lucas, J. R. (1970). *The concept of probability*. Oxford, UK: Clarendon Press.
- Mellor, D. H. (1971). *The matter of chance*. Cambridge, UK: Cambridge University Press.
- Mises, R. v. (1957). *Probability, statistics and truth* (2nd English ed.). New York, NY: Macmillan. (Original work published 1928).
- Nagel, E. (1945). Is the Laplacean theory of probability tenable? *Philosophy and Phenomenological Research*, 6, 614–618.
- Partridge, E. (1966). *Origins* (4th ed.). New York, NY: Macmillan.
- Peikoff, L. (1979). The analytic-synthetic dichotomy. In A. Rand (Ed.), *Introduction to Objectivist epistemology* (pp. 117–164). New York, NY: New American Library. (Original work published 1967).
- Peirce, C. S. (1878a). The doctrine of chances. *Popular Science Monthly*, 12, 604–615.
- Peirce, C. S. (1878b). The probability of induction. *Popular Science Monthly*, 12, 705–718.
- Pepper, S. C. (1942). *World hypotheses*. Berkeley, CA: University of California Press.
- Popper, K. R. (1968). *The logic of scientific discovery* (2nd English ed.). New York, NY: Harper Torchbooks. (Original work published 1934).
- Quine, W. (1951). Two dogmas of empiricism. *Philosophical Review*, 55, 20–41.
- Reichenbach, H. (1949). *The theory of probability* (2nd ed.). Berkeley, CA: University of California Press.
- Spencer Brown, G. (1957). *Probability and scientific inference*. London, UK: Longmans, Green.
- Venn, J. (1888). *The logic of chance* (3rd ed.). London, UK: Macmillan. (1st ed., 1866).
- Werner, H. (1937). Process and achievement—A basic problem of education and developmental psychology. *Harvard Educational Review*, 7, 353–368.
- Williams, D. (1945). The problem of probability. *Philosophy and Phenomenological Research*, 6, 619–622.
- Yeo, R. R. (1986). Scientific method and the rhetoric of science in Britain, 1830–1917. In J. A. Schuster & R. R. Yeo (Eds.), *The politics and rhetoric of scientific method: Historical studies* (pp. 259–297). Dordrecht, The Netherlands: Reidel.

Chapter 7

The Fisher and Neyman-Pearson Theories of Statistical Inference



7.1 Fisher

The orthodox theory of statistical inference is due to Jerzy Neyman and Egon S. Pearson, whose work is a reformulation of the Fisherian corpus. Tempting as it was to title this chapter simply “The Frequentist Theory of Statistical Inference,” elegance and economy would have been served more than accuracy. Though Fisher took a frequency interpretation of probability as a design criterion in his theory of statistical inference, he also equivocated on it all his life and argued hotly with Neyman and Pearson against some of the central implications of the frequency interpretation. Consequently I shall consider their approaches separately, before undertaking a comparison at the end of the chapter.

Fisher, indisputably the greatest statistician of all time, developed most of the statistical techniques used by psychologists today and was the first to give systematic attention to the logic of modern statistical inference. His output was prodigious: Apart from his 6 books and hundreds of published letters and reviews, his bibliography includes 294 articles in some 90 journals.

He was interested in both mathematics and biology from an early age; Yates and Mather (1963) report that his mind was made up one day when “on a chance visit to a museum he happened on a cod’s skull with all its bones separated and labelled with their names; he decided on mathematics” (p. 92). Partly through Karl Pearson’s influence, he developed a lifelong interest in genetics and evolution; his contributions in this field alone would have been enough to secure him an international reputation. Indeed, in 1911, he gave a short talk to the Cambridge University Eugenics Society which reconciled the hostile schools of Mendelism and biometry—when he was 21 years old. Apart from his theoretical contributions to the field, he also played a major role in understanding the genetics of human blood groups.

His interest in statistics dated from his student days, when he read Pearson’s biometrics papers of the 1890s in the *Philosophical Transactions* and also kept up

with *Biometrika*. When several colleagues of Pearson's published a progress report of their work on the distribution of the correlation coefficient in small samples, Fisher was inspired to work out the full solution in closed form; he allegedly sent in a rough draft of the paper within a week (Mahalanobis, 1938). The problem had eluded Pearson's efforts for years; his acceptance of the paper for publication in *Biometrika* (Fisher, 1915) was, according to Neyman (1967b), one of the two occasions when Fisher and Pearson were known to have agreed on anything. Fisher was 25 at the time.

His solution was revolutionary in its approach: Whereas we are accustomed to representing a correlation as a plot of n points in two-dimensional space, Fisher realized that it could equally well (in principle, if not in practice!) be represented as two points—an X point and a Y point—in n -dimensional space, one dimension for each individual. The correlation then becomes the cosine of the angle between the vectors drawn from the origin to the points representing the two variables. At 90°, the cosine is 0; if the two vectors coincide, the angle is 0°, and the cosine is 1. That insight, with transformation to polar coordinates, led (eventually) to the solution. Fisher's amazing ability to visualize complex geometrical relationships is commonly attributed to his poor eyesight (cf., e.g., Mahalanobis, 1938); his mathematics tutor had him solve problems without pencil and paper. Kendall (1963) is skeptical of this explanation, but, whatever its origins, the ability put Fisher's work in distribution theory on a plane which few of his contemporaries could understand. In the papers that followed, he obtained the distributions of the single and multiple regression coefficients (1922a), as well as partial regression coefficients (1924), and proved the argument given by Gosset for the t distribution (1925a). The first analysis of variance was published in this period (Fisher & Mackenzie, 1923), and the analysis of covariance soon followed (Fisher, 1925b/1932). His work largely inaugurated, in addition, the whole field of experimental design. Kempthorne (1976) would seem to be entirely justified in describing Fisher as "about one Gauss in stature" (p. 304).

Despite his unparalleled achievements in the field, recognition was slow to come, and Fisher for a long time had trouble getting his work published. Part of the problem was the unexampled difficulty of his work; a number of Fisher's colleagues have commented that, when they reached one of his favorite expressions, "It is therefore obvious . . .," they knew they were in for a good many hours of heavy algebra before they got to the next line. Another part of the explanation lies in Fisher's notorious arrogance and insensitivity. His daughter and biographer, Joan Fisher Box (1978), suggests a connection with his mother's emotional remoteness; of Fisher, she says, "He was at once exceedingly self-centered and utterly self-forgetful, charming and impossible" (Box, 1978, p. 12).¹

Though Fisher came to the field with a strong predisposition for interpersonal difficulties, historical circumstances were also such as to bring out the worst. The

¹Evidently he also fit very well the model of the absentminded professor. Box's book is full of wonderful anecdotes about Fisher appearing at a formal evening function in bedroom slippers, or, as he gesticulated in a fit of rage, crushing to death a mouse he was holding in his hand,

irascibility and contempt he was eventually famous for were even more pronounced in his predecessor Karl Pearson, and one gets the impression of Pearson's quirks reverberating through successive generations of statisticians in imitation of intrafamilial transmission of interpersonal dynamics. Pearson himself had endured in turn some publishing frustrations at the hands of William Bateson, which had led to the founding of *Biometrika*.

Pearson had been big enough to publish the upstart Fisher's paper on the distribution of the correlation coefficient in 1915, but when Fisher submitted a major paper the next year on Mendelism and biometry to the Royal Society, Pearson the biometrician and Punnett the geneticist both rejected it as referees; Leonard Darwin, a son of Charles, eventually arranged for its publication by the Royal Society of Edinburgh. In 1919 Pearson offered Fisher a position at the Galton Laboratory; both men were no doubt aware of the implications of Fisher's working under Pearson; and Fisher, having also received an offer at Rothamsted Experimental Station, accepted the latter. In 1920 he submitted another paper on the correlation coefficient to *Biometrika*; Pearson refused it, as did the Royal Statistical Society (on the grounds that it was too mathematical), and Fisher resorted to a new Italian journal, *Metron*.

The Statistical Society did publish a major paper of his in 1922 on the χ^2 test for an $R \times C$ contingency table, showing that the degrees of freedom were $(R - 1)(C - 1)$ when expected frequencies were calculated from the marginal totals. Since

the key mouse, a male whose progeny should have clinched the genetical argument. That night he could settle down to nothing, was restless, wretched beyond anger. He picked up Descartes and then resorted to Marcus Aurelius, whose stoicism, if anything, might fortify his own to accept what he had done. (Box, 1978, p. 170)

The story of his 1953 visit to Australia appears to have been typical. On his arrival he had left his coat on the plane, and he lost the return half of his ticket back to London. Then:

The departure had not gone smoothly. They had to make an early start, about 5 a.m. The Wolseley was out of order and Cornish used the 1928 Dodge. The car was loaded and ready to go when Fisher asked, "Where's my bag?" It took Cornish a few minutes to discover it, where Fisher had laid it down outside the front door (he had gone out that way first and the door had shut behind him; leaving the bag, he had then made his way through the undergrowth to the back door by the garage to pick up hat, coat, and stick). Backing out in the darkness, the car ran over a child's scooter in the driveway which jammed up over the fan blade. They extricated the scooter and drove off in a hurry. The fan blade had been bent, however, and, by a tin-opener effect, had sliced open the bottom end of the radiator, so that the car broke down halfway to the railway station. Fisher got out of the car, carrying his case. Cornish, now anxious about time, instructed him in firm tones: "Wait here for me. Just stay here while I go and get a taxi." When he returned, of course, Fisher was nowhere to be seen. As they transferred the luggage to the taxi, the cabbie observed "an old codger with glasses" coming toward them, who hailed them: "Ah, Alf, you found a taxi." But where was his brief case? It was too late now to ask. They rushed to the station, brought out tickets, paid off the taximan; Cornish found Fisher halted talking with a porter, hauled him off, put him aboard, and told the conductor, with relief, "He's all yours!" (pp. 468–469)

Pearson's introduction of the test in 1900, the degrees of freedom had always been taken to be one less than the number of cells, and discrepancies had begun turning up in empirical work. (For a detailed discussion of this controversy and its implications, see Baird, 1981/1982.) The χ^2 test was Pearson's crowning achievement, and Fisher had shown it to be in error in one of its major applications. Pearson's pride was severely wounded, and he lashed out with a contemptuous attack in *Biometrika*, rejecting Fisher's argument. Fisher submitted a rebuttal to the Statistical Society, which, evidently intimidated by the dispute between the two giants, declined to publish it. Fisher resigned the Society in protest, but his reputation suffered in consequence. The final coup came in 1926, when Pearson's son Egon published a report in *Biometrika* on an empirical investigation of Bayes' Theorem. Box (1978) writes:

These results comprised some 12,000 fourfold tables observed under approximately random sampling conditions. From them Fisher calculated the actual average value of χ^2 which he had proved earlier should theoretically be unity and which Pearson still maintained should be 3. In every case the average was close to unity, in no case near to 3. Indeed, Fisher pointed out that the general average, with an expected value of 1.02941 ± 0.01276 , was "embarrassingly close to unity" at 1.00001, and from this he inferred that the sampling conditions had not been exactly random. There was no reply. (p. 88)

The antagonism between Pearson and Fisher persisted until the former's death, or beyond. Although Pearson had not invented the method of moments, he had spent much of his life fitting frequency curves by this means. In 1935 R. S. Koshal published an article demonstrating the feasibility of the method of maximum likelihood for this purpose, rousing Pearson to a vitriolic attack. Fisher published a defense of Koshal, whose work had borrowed from his own, and in response Pearson printed an extremely sarcastic and hostile response in *Biometrika*. It was his last publication; he died in 1936. Fisher then published a final response; Box writes that he "felt free . . . to consider in all seriousness the question Pearson had raised" (p. 330); Neyman, on the other hand, found shockingly indecorous Fisher's attack on the recently deceased (Reid, 1982).

When Pearson retired in 1933, Fisher was the logical choice to succeed him in the Galton Chair of Eugenics. Pearson's son Egon had been working in the Galton Laboratory for 10 years, however, and was accordingly a natural choice on other grounds. The dilemma was resolved by creating a new Department of Applied Statistics at University College, London, for Egon to head. Fisher felt perfectly qualified for both positions and regarded the younger Pearson as a nonentity (Reid, 1982). One gets the impression that Pearson did not wholly disagree; he was, in any event, understandably intimidated by Fisher. Pearson approached Fisher at the beginning with the request that Fisher not lecture on statistics, so as to minimize conflict with his department, which comprised mostly the older Pearson's staff. Fisher partially acceded, but the fact that these two rather antagonistic departments were housed on adjacent floors of the same building and shared the same common room for tea produced some "Gilbertian situations" (Yates & Mather, 1963).

In the summer of 1931 and again in 1936, Fisher visited the Iowa State College of Agriculture in Ames, at the invitation of George Snedecor. Snedecor, a member of the mathematics department, was one of the first Americans to take an interest in

Fisher's work; he and Henry Wallace (later Secretary of Agriculture and Vice President under Roosevelt) helped to found a statistical computing center which was the first of its kind in the United States. The second visit was the occasion also for Fisher's receiving an honorary degree from Harvard at its tercentennial celebration and for his being offered a professorship at Ames. His unhappiness with both his professional and domestic situation² made the offer tempting; had it not been for the summer weather in Ames, the subsequent character of American statistics might have been somewhat different: Staying at a campus fraternity house, Fisher coped with the heat only by bringing his bedsheets down every morning and stuffing them in the refrigerator for the day (Box, 1978).

In 1943, Fisher accepted the Arthur Balfour Chair of Genetics at Cambridge University, which he held until his retirement in 1957. In this position he succeeded Karl Pearson and Reginald Punnett, the two referees who had rejected his paper on biometry and Mendelism in 1916. It is remarkable, as Kendall (1963) notes, that the world's leading statistician never held a chair in statistics.

7.2 The Fisherian Theory of Statistical Inference

Intellectually, Fisher was solidly in the tradition of thinkers like Venn, who were very much concerned with the philosophy of knowledge, but with a decidedly British pragmatism and common sense. Like Venn, he was also concerned to avoid the difficulties of Laplacean inverse probability. His approach to statistical inference was guided, like everyone else's, by his views on probability. But, as Kyburg Jr. (1974) says, "It would be unfair to criticize 'Fisher's interpretation of probability,' because he never actually gives one" (p. 74). When he did speak of probability, his conception was rather close to Venn's. He not only insisted on a frequency definition, but also wanted to keep the concept of probability closely linked with the problem of inference, so that his statements about probability characteristically contain reference to knowledge and ignorance. The tension between "aleatory" and "epistemic" aspects of probability reaches perhaps its greatest height in Fisher's work. It was his rejection of an epistemic or subjective concept, in terms of degrees of evidence or belief, that led to his formulation of a frequentist theory of statistical

²Fisher, having fathered eight children, is credited uniquely among eugenicists for having practiced what he preached. He was, in general, more socially conservative than revolutionary; MacKenzie (1981) notes: "It is difficult to avoid the impression that amongst the social interests sustaining 'positive eugenics' may have been that of professional men in having women--especially the growing number of women professionals--return to their traditional roles and stop 'shirking' motherhood" (p. 42). He was, at all events, too busy professionally to be very involved with his family and, given his personality, would have been less than an optimally sensitive husband. The burden on Eileen Fisher, raising eight children on a farm near Rothamsted, with a modest income, was heavy. They split when Fisher went to Cambridge to accept the Balfour Chair in 1943, and he seems to have been little involved with his family after that. When he died in Australia in 1962, he was thousands of miles from all of them, and had left no will.

inference in the first place; yet it was his insistence that probability was still an epistemic phenomenon that gave his theory its ostensible relevance in scientific research. His attempt essentially to have his frequency cake and eat it, too, has not generally been viewed as a success, and whatever criticisms his ideas have received spring from just this tension.

Fisher's name is prominently associated with three devices of statistical inference: maximum likelihood, significance testing, and fiducial probability. If none of these was purely his invention *de novo*, and if the one which was most distinctively his own was also the least successful (fiducial probability), he is still entitled far more than anyone else to the credit for the widespread use of statistical inference today.

Under the frequency theory of probability, the object—and the challenge—of statistical inference is to make statements about a population or hypothesis or cause or set of generating conditions on the basis of some observations, when it is the observations rather than the model that have a probability distribution. Significance testing, the device most familiar to psychologists, achieves that object by substituting decision for inference—though Fisher equivocated on this point. Maximum likelihood selects out those values for the population parameter in question which would make the sample results most likely. Fiducial probability, the most controversial procedure, attempts to turn the probability distribution around to apply to the parameters rather than the statistics.

7.2.1 Maximum Likelihood

The method of maximum likelihood was the subject of Fisher's first publication (1912). The basic idea had been suggested by Daniel Bernoulli in 1777 (Kendall, 1961) and by J. H. Lambert even before that, in 1760 (Sheynin, 1971); Fisher's contribution was to dignify it as a general method and to name the forward probability of the results their likelihood. Likelihood theory is involved behind the scenes, as it were, in many textbook procedures,³ but has not been prominent as a method of inference in its own right until recent work by A. W. F. Edwards (1972) and

³In a one-way analysis of variance, for example, the maximum likelihood justification proceeds by forming an expression for the joint probability density functions of the observations over the entire parameter space Ω and over the subset ω of that space designated by the null hypothesis. These joint density functions are the likelihood functions $L(\Omega)$ and $L(\omega)$ of the sample, under the alternative and null hypotheses, and they may take on a whole range of values. The method of maximum likelihood arbitrarily chooses the maximum values assumed individually by $L(\omega)$ over ω and by $L(\Omega)$ over Ω and then takes their ratio, λ . In analysis of variance, the method leads to a ratio of within-group to total variance. Small values of λ imply a departure from null expectation and thus afford a criterion of significance. In multivariate analysis of variance, this likelihood ratio is itself one of the commonly used criteria for testing; in the univariate case, an inverse monotonic transformation to the familiar F yields a test with identical results.

others (Chap. 10). It serves here, however, as a brief but useful illustration of how a theory of statistical inference can be got out of a frequentist theory of probability.

Since the frequency conception of probability constrains us to speak only of forward probabilities—probabilities of outcomes, given certain conditions as true—the method of maximum likelihood simply takes as the best estimate of the parameter that value which maximizes the probability of the results obtained. The principle of maximum likelihood holds essentially that whatever happens is always the most likely thing that could have happened. As a substantive principle, it is obviously false; but as a heuristic principle—its intended use—it has plausibility. It presents potentially two subtle difficulties for the frequency conception.

The first lies in the derivation of the maximum likelihood value. In that process it is necessary to construct a parameter space and to integrate over it and thereby to treat the unknown parameter as a variable rather than a fixed constant. Karl Pearson jumped on just this point in criticizing Fisher's (1912) paper, and Fisher was later (1922b) apologetic about the implicit use of inverse probability, agreeing that it was indeed improper to speak of the probability of a parameter θ lying within a given interval. In a later paper he attempted a further clarification:

The function of the θ 's maximised is not however a probability and does not obey the laws of probability; it involves no differential element $d\theta_1 d\theta_2 d\theta_3 \dots$; it does none the less afford a rational basis for preferring some values of θ , or combination of values of the θ 's, to others. It is, just as much as a probability, a numerical measure of rational belief, and for that reason is called the *likelihood* of $\theta_1, \theta_2, \theta_3, \dots$ having given values, to distinguish it from the probability that $\theta_1, \theta_2, \theta_3, \dots$ lie within assigned limits, since in common speech both terms are loosely used to cover both types of logical situation. (Fisher, 1930, p. 532)

It is indeed possible to regard h as a variable without regarding it as a random variable, which would entail a probability distribution; this is what is done in plotting an ordinary power curve through the “parameter space.” But it is not at all clear how the likelihood might be maximized without differentiation with respect to the θ 's.

He went on to say:

Knowing the population we can express our incomplete knowledge of, or expectation of, the sample in terms of probability; knowing the sample we can express our incomplete knowledge of the population in terms of likelihood. We can state the relative likelihood that an unknown correlation is +0.6, but not the probability that it lies in the range .595–.605. (1930, p. 532)

This passage points up the second difficulty, which is semantic, and peculiarly Fisherian. Since the whole point was to compare hypotheses—possible values of the parameter θ —with regard to how well supported they were by the data, Fisher seized on the equivocation contained in the word *likelihood*:

I suggest that we may speak without confusion of the *likelihood* of one value p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample. (1922b, p. 326)

Fisher's statement here, however, amounts to a tacit inversion, suggesting that he was a bit glib in assuming no confusion would ensue. Likelihood as defined is a

relative frequency probability of observations on some hypothesis; yet he is speaking of the likelihood of some value of the parameter p being thrice another. Hence it cannot, as he acknowledges, be a synonym for probability. But in spite of this caveat, Mises (1928/1957) complains that:

As soon as he has introduced his new term, Fisher starts using it as completely synonymous with “probability” in the everyday sense of the word. That is, he considers a value of p as more reliable if it has a greater likelihood, and he recommends in particular, as the “best estimate” of p , the value which has the greatest likelihood. (p. 158)

Likelihood theory does not necessitate the Fisherian locutions, such as speaking of one hypothesis as three times as likely as another; that is merely a temptation into which Fisher, being first in line, fell. Other formulations of likelihood theory will be presented in Chap. 10.

7.2.2 *Significance Testing*

In the early years of the twentieth century, the idea of significance testing, as had been glimpsed in the work of Edgeworth and Venn, had not caught on; the focal problem of statistics, as noted earlier, was estimation. When Fisher published his first major paper “On the Mathematical Foundations of Theoretical Statistics” in the *Philosophical Transactions* (1922b), he listed under the heading “The Problems of Statistics” “(1) Problems of Specification” (“the choice of the mathematical form of the population”), “(2) Problems of Estimation,” and “(3) Problems of Distribution” (“the distribution of statistics derived from samples”), but significance testing does not appear. It was to evolve first as a kind of substitute for estimation, later to overtake it as a strategy of inference in its own right.

Small-Sample Theory Fisher’s most distinctive contribution to significance testing was that contained in the shift from large-sample to small-sample methods, and it was this shift that made the new theory specifically a theory of *statistical* inference. Yule (1911) give some indication how the switch came about.

It cannot be overemphasized that estimates from small samples are of little value in indicating the true value of the parameter which is estimated. Some estimates are better than others, but no estimate is very reliable. In the present state of our knowledge this is particularly true of samples from universes which are suspected not to be normal.

Nevertheless, circumstances can sometimes drive us to base inferences, however tentatively, on scanty data. In such cases we can rarely, if ever, make any confident attempt at locating the value of a parameter within serviceably narrow limits. For this reason we are usually concerned, in the theory of small samples, not with estimating the actual value of a parameter, but in ascertaining whether observed values can have arisen by sampling fluctuations from some value given in advance. For example, if a sample of ten gives a correlation coefficient of +0.1, we shall inquire, not the value of the correlation in the parent universe, but, more generally, whether this value can have arisen from an uncorrelated universe, i.e., whether it is *significant* of correlation in the parent. (p. 437)

Thus significance tests, instead of yielding descriptive statements estimating population parameters, provided yes-or-no answers to questions, which, in comparative agricultural (or medical) experiments, typically took the form, “Did the treatment have an effect?” Here, what would have been regarded as a limitation was turned around to an advantage. Large-sample statistical methods could never yield a simple answer to such a question; they could only give an answer, however precise, to the question of *how large* a difference was produced by the treatment over the control. But the question of the *significance* or *importance* of this difference was left up to the judgment of the experimenter. In small-sample theory, on the other hand, once the 5% level was accepted as the rule, experimenters had, mirabile dictu, an unambiguous criterion for determining whether a difference was significant, or, for short, whether there *was* an effect due to treatment. As was noted in Edgeworth’s work, the word *significant* was perfectly chosen to reflect both the property of being indicative and of being important, which was more the experimenters’ concern. Hence it was that small-sample theory, far from being a poor substitute for the method of large samples, went well beyond it toward a theory of statistical inference *proprement dite*.

The shift from large-sample to small-sample methods coincided with a changing emphasis from estimation to significance testing. In large-sample theory the aim was a summary description of some natural population, and the method was to take a sufficiently large sample of the population to afford a reasonably realistic representation in the sample. The sample was still conceived to be random, and that assumption was necessary in the derivation of sampling distributions for calculating probable errors; the sample value could be counted on to be acceptably close, in most instances, to the population parameter. In small-sample theory, on the other hand, the feature of randomness came to the fore. Individual samples of 10 or 20 observations were so variable, as Yule (1911) note, that the estimates they gave could not be relied upon. Thus it is here that the method becomes truly statistical: We abandon statements at the individual level and confine our predictions to the aggregate—the aggregate comprised by the sampling distribution. Theoretically, we are no more entitled to make anything of an individual outcome in random sampling than in tossing pennies; we say only, for example, that we expect no more than 5% of z s to exceed 1.96 in absolute value.

The distinction between large-sample and small-sample methods, which was so prominent in statistics texts before World War II, is largely unfamiliar to present-day psychologists, as large-sample methods have virtually disappeared in this field (except in psychometrics); it survives only in the distinction between z and t tests, and even there many texts fail to capture adequately the sense of the distinction. Fundamentally, the difference is only one of degree, if we think of sample size as a continuous variable. But the difference becomes nearly categorical, in the distinction between heuristics of representativeness and randomness, which will be discussed in Chap. 10.

The concept of significance testing with small samples could not have succeeded so easily had not some conventional criteria been established rather quickly. The rivalry between Fisher and Karl Pearson played an odd role here. The 5% tail

integral derived, as was described in Chap. 5, from Chauvenet's procedure for the rejection of "doubtful" measurements in astronomical work, where the basis of doubt was sheerly the magnitude of the value, and the criterion for rejection was strictly probabilistic. It had become customary, owing to the typical number of observations made of astronomical phenomena, to regard an observation greater than twice the standard error, or three times the probable error, as a clear candidate for rejection. In a normal distribution, two times the standard error cuts off roughly the extreme 5%. When Fisher published his *Statistical Methods for Research Workers* in 1925, he proposed to define the criterion directly in terms of probabilities rather than probable errors or standard errors, and suggested that the 5% and 1% levels be adopted as a convention, as a matter of convenience. He appealed to Pearson for permission to reprint in *Statistical Methods for Research Workers* some of the tables from Pearson's *Tables for Biometricalians and Statisticians*. Pearson refused, at least partly because he feared the effect on sales of his own *Tables*, on which he depended for support for *Biometrika*, and Fisher was forced to prepare his own tables. Departing from the established practice, which gave probability values for points all across the distribution, Fisher organized his tables by quantiles. In providing extensive tables at just the .05 and .01 levels, Fisher facilitated thinking of the cut-off as a function of the associated probability, rather than vice versa. That decision, as much as any other single event, shaped the subsequent practice of statistical analysis. Fisher and Yates followed the same procedure when they published their *Statistical Tables for Biological, Agricultural and Medical Research* (1938). With only a few irregularities, new editions of *Statistical Methods for Research Workers* were published every 2 years from 1928 to 1950; during those years, permission to reprint tables from one or the other of these books was granted to over 200 authors of statistics texts (Box, 1978).⁴

The Hypothetical Infinite Population Small-sample theory brought into focus the distinction between sample and population; it is not surprising that Student (1908) should have been the first to introduce notation to mark the distinction. Paradoxically, however, in the agricultural work in which he and Fisher were involved, neither the sample nor the population had any reality. There was no population of yields of barley from which the present data had been randomly sampled. Fisher and Gosset were simply looking for a way to make sense of a small body of observations resulting from a single experiment. Fisher solved the conceptual

⁴ Nowadays computers can generate new tables with an effort that is hardly noticed, but it is important to remember in this history what an arduous tedium the compilation of statistical tables was a century ago. Pearson's staff spent years working on his tables, with calculators that were slow and primitive. Gosset described in a letter to Fisher the trials of his assistant E. M. Somerfield in producing a table of the *t* distribution:

Yesterday I found him with the machine which Noah used when quantity surveying before his voyage. The story goes that he subsequently bartered it for a barrel of porter with the original Guinness. Anyhow he doesn't seem to have been able to keep it dry and Somerfield wasn't strong enough to turn the handle. (quoted in Box, 1978, p. 117)

problem in his “Mathematical Foundations” paper (1922b), with his enormously important notion of the hypothetical infinite population. He introduced it as follows:

Briefly, and in its most concrete form, the object of statistical methods is the reduction of data. . . .

This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. Any information given by the sample, which is of use in estimating the values of these parameters, is relevant information. Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by any actual sample is irrelevant. It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data. (1922b, pp. 311–312)

It should be noted that there is no falsehood in interpreting any set of independent measurements as a random sample from an infinite population; for any such set of numbers are a random sample from the totality of numbers produced by the same matrix of causal conditions: the hypothetical population which we are studying is an aspect of the totality of the effects of these conditions, of whatever nature they may be. The postulate of randomness thus resolves itself into the question, “Of what population is this a random sample?” which must frequently be asked by every practical statistician. (p. 313)

The hypothetical population is regarded as infinite for two reasons: (a) It allows probabilities to be extended to the real line, including irrational as well as rational numbers; and (b) in a finite population, knowledge of one element affects probability statements about others, so that sampling probabilities do not remain constant; and this, Fisher (1925c) says, “is not the hypothesis which we wish to consider” (p. 700). The infinity of values in the hypothetical population generates in turn an infinity of values for the sampling distribution.

The present generation of psychologists, schooled in the Fisherian doctrine, has almost succeeded in making sense of the concept of the hypothetical infinite population, but to Fisher’s contemporaries, and to philosophers down to the present, the concept offers some serious problems. Kendall (1943) confesses his difficulty with the notion of regarding our data as a random sample from a hypothetical infinite population:

There are obvious logical difficulties in regarding such a sample as a selection—it is a selection without a choice—and still greater difficulties about supposing the selection to be random; for to do so we must try to imagine that all the other members of the population, themselves imaginary, had an equal probability of assuming the mantle of reality, and that in some way the actual event was chosen to do so. This is, to me at all events, a most baffling conception. (p. 187)

The two most serious problems with the hypothetical infinite population are its vagueness and its circularity. Speaking directly to Fisher’s conception of the inference process, Kempthorne, a former associate of Fisher’s, says:

One cannot define a real population from a sample, and the relevant populations cannot often be given a specification; we do not know what would happen in repeated trials because

we do not know how to repeat the trial; i.e., we do not know how to get a random sample from “the population from which the present experimental units are a random sample.” (Kempthorne & Doerfler, 1969, p. 235)

In general, in inventing a fictional population for the purpose of inferring properties of the individual case before us, we are skirting circularity, if not actually falling right into its lap. When we “imagine a population round a given sample of field trials and then use those trials to infer the properties of [our] imaginary population,” Kendall (1942) wonders “whether, in fact, in Eddington’s expressive adaptation of Defoe, the footprints we keep discovering in the sand may not be our own” (p. 72).

It would seem possible that the theory of significance testing could have developed in a less circuitous way: If the subject matter in some field limited investigations to small samples, the variability of such samples would have indeed called attention to the stochastic nature of the sampling process, and to the futility of estimation, and significance testing might have evolved as a kind of poor substitute in just the way Yule (1911) suggest. Interestingly, so-called small-sample theory arose in a context which had nothing to do with sampling. A small body of observations had been made, but there was no very clear reference to a larger population. It was for the purpose of bestowing meaning on these limited results that the hypothetical infinite population was invented, on the model of the real populations studied by the biometricalists. If its vagueness and circularity occasioned little concern among practitioners, it was because there was no real population of interest, and the whole object was a statement of the significance or importance of the few data at hand.

Randomization Tests In view of the difficulties with the concept of the hypothetical infinite population, it is interesting that Fisher tended to invoke it even where it was unnecessary. As Hacking (1965) observes:

A procedure for sampling a population certainly involves a chance set-up. But only excessive metaphor makes outcomes of every chance set-up into samples from an hypothetical population. One hopes our logic need not explicitly admit an hypothetical infinite population of tosses with this coin, of which my last ten tosses form a sample. (p. 25)

Hacking’s point is relevant to the most famous of all Fisher’s examples. Shortly after moving to Rothamsted, he was about to serve tea one day to Muriel Bristol, an algologist, who objected to Fisher’s not having put the milk in the cup first. Wondering whether she could really tell the difference, someone proposed a test. The details of the experiment were not recorded, but Bristol evidently acquitted herself admirably (Box, 1978). Years later, Fisher (1935a), though he subsequently claimed not to remember the inspiration, used this example to open the second chapter of *The Design of Experiments*. The anonymous lady was given eight cups, four with the milk having been added first and four with the tea infusion first. The probability of her guessing all 8 correctly by chance would be $4!4!/8!$, or 1/70. On a frequency interpretation of probability, this would mean that in infinitely many repetitions of the experiment, all 8 should be correctly guessed in only 1/70 of them.

It is obviously necessary to assume that the conditions of the experiment don't change: The lady doesn't learn from experience, her taste buds don't wear out, etc.

As Hogben (1957) points out, it is not clear that the services of a statistician would be needed: If she can discriminate, one or two repetitions would suffice to persuade any skeptic; if she cannot, the matter is settled even more easily. Fisher was thinking of a lady who could not make the distinction invariably, but who would be right more often than not. In this case we may think of her sensory apparatus as a sort of chance machine, which we are testing for bias. To regard these eight guesses as a random sample from a hypothetical infinite population of teacup guesses involves a further ambiguity, however. Hogben quotes Raymond Wrighton:

We may identify *the lady as she is at the time of the experiment* with a conceptual lady out of whose infinitely protracted tea-tasting the experience of the experiment is regarded as a random sample. The idea may be attractive, but it carries with it an embarrassing consequence. . . . If the experiment demonstrates the phenomenon, it is the conceptual lady who must in fairness be rewarded, and if not, it is the conceptual lady whose pretensions must be exposed. (1957, p. 471)

The implicit intrusion of the concept of sampling from a hypothetical infinite population created another embarrassment in Fisher's introduction of the exact randomization test for two independent groups. He proposed to test the difference in mean height between a sample of 100 Englishmen and a sample of 100 Frenchmen, drawn from their respective populations, by inscribing all the heights on slips of paper, shuffling them, and separating them at random into two groups of 100. The reference class for the significance level probability is then the set of all possible such divisions. But it is not clear where the logical warrant for this operation lies. The procedure appears to imply that we could take a "man without a country," randomly assign him to a nationality "treatment," and observe the effect on his height. Here the experimental "randomization" is hypothetical, just as the random sampling was before. We have to imagine that nature, as it were, randomly assigned individuals to each country of birth, and then ask whether the observed difference in height is more unlikely than in most of the other random assignments that might have occurred. But just how meaningful is the question? Does a randomization test make sense when applied to sex differences? It is not clear, once we allow random assignment to be hypothetical, how any lines can be drawn in principle. If we want to exclude any cases at all as ridiculous, such as hypothetically assigning neuter individuals a sex and observing its differential effects on them, we shall evidently have to do so on more or less ad hoc grounds.

I notice from Anthony Oberschall (1987) that Maurice Halbwachs was making much the same point over 80 years ago. Commenting on the statistical practice of comparing mortality rates between countries while adjusting for differences in the ages of the respective populations, he asked:

How long would the French live if, remaining French, they lived in the same physical and social conditions as the Swedes? How long would the Germans live if, remaining German, they lived in the same conditions as the French? That amounts . . . to wondering how long a camel would live, if, remaining a camel, it was transported to the polar regions, and how long a reindeer would live if, remaining a reindeer, it was transported to the Sahara. In other

words, everything is done as if, in order to study the demographic characteristics of a country, it was necessary to start from a population which is not that of any country, as if one had to work with men who weren't born, didn't marry, didn't die in any definite region in any particular manner, with respect to familial, religious, juridical, economic customs. But, just like the homo economicus, such a homo demographicus is an abstraction too carefully detached from reality to tell us anything about what is real. (Halbwachs, 1935/1972, pp. 337–338)

Fisher argued, nevertheless, that randomization provided a logical basis for sampling-theory tests, so that the conclusions in, say, a conventional t test “have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method” (Fisher, 1936a, p. 59). Earlier, in introducing the exact randomization test for the matched pairs design, he said: “It seems to have escaped recognition” in discussions of assumption violation,

that the physical act of randomization, which . . . is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. (Fisher, 1935a, p. 51)

Modern psychologists have often followed Fisher in treating randomization tests as ordinary distribution-free alternatives to standard normal-theory tests; but, in logic, the choice of tests is dictated by what the experimenter in fact did—random sampling or random assignment—and the nature of the inference in the two models is correspondingly different. Whereas the familiar sampling theory licenses a “statistical inference” about the putative infinite population of interest, any inference about individuals other than the experimental units in the randomization model must be made on extrastatistical grounds; its reference is to the “population” of possible assignments of individuals to conditions. In practice, of course, the model of random assignment is much easier to justify in most laboratory research. And Kempthorne, the most distinguished advocate of randomization tests, makes the interesting point that:

The population of repetitions involved in the randomized experiment, while conceptual after the performance of the experiment, is entirely real before the performance of the experiment, and may be contrasted favourably as regards vagueness with the population of repeated samples from a population of experimental units. (Kempthorne & Doerfler, 1969, p. 239)

All significance testing involves hypothetical suppositions, reference to what would be the case if some events occurred which did not in fact occur. We do not ordinarily balk at the idea that a person *could have* turned up in another group, provided that we have randomly assigned people to groups in the first place; it is much less clear what it means to say that a given Englishman *could have* been French, much less yet to say what the effect would have been on his height. Somewhere here we are drawing distinctions between what we are and are not willing to ascribe to the mechanism of chance, between what could and what could not have happened differently from the way it did.

The general question underlying the probabilistic evaluation of experimental results, of course, is “What could have happened differently?” The rationale of

significance testing, however, licenses only a very narrow range of answers: We would either have drawn a different sample or made a different assignment at random, depending on which we did in the first place. Use of the word *chance* in this context is hazardous, just because it potentially covers a much larger range of factors (e.g., it was only chance that individual X is English or female). It is important to recognize that when significance tests are said to control for the chance hypothesis, or to solve the problem of great variability in psychological data, it is only the variability due to sampling or assignment that is properly meant. These are not the only sources of variability, and they may not be the most important. Enlow (1937) was making the same point years ago, before the practice of significance testing became widespread:

Obviously, in measuring the achievement (say) of a given class, one is more interested in the likely extent of mismeasurement due to response error of the pupils measured than in the probable deviation of the class from a fictitious “universe,” due solely to the chance fluctuation of pretended unbiased sampling. (p. 130)

The Success of Significance Testing As is sometimes true in the evaluation of other historical change, whether the growth of significance testing was rapid or glacial depends on one’s viewpoint. Thus, while Hotelling (1951) remarks on how long it took for Fisher’s methods to be accepted into university curricula, Welch (1968) is equally impressed with how quickly it took place. The practice did not become common in psychological research until the 1950s, but it is most interesting to see, even in the early years of its application in agricultural research, the extraordinary pull of significance testing for its own misuse. The criticisms leveled by some of those closest to Fisher show clearly the tendency, from the beginning, for the question of significance to override all others and for the stars appended to statistical results to be treated somewhat like those awarded for perfect attendance or military exploits.

Frank Yates, a colleague of Fisher’s and perhaps his stanchest defender, blamed the problem partly on the “excessive emphasis on tests of significance” (Yates & Mather, 1963, p. 113), which he considered the principal weakness of *Statistical Methods* and *The Design of Experiments*.

In the first place it has resulted in what seems to me to be an undue concentration of effort by mathematical statisticians on investigations of tests of significance applicable to problems which are of little or no practical importance. Second, and more important, it has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating.

Historically, this situation is understandable. When *Statistical Methods* was written the methods used for testing significance were . . . in the utmost confusion. In the interpretation of their results research workers in particular badly needed the convenience and the discipline afforded by reliable and easily applied tests of significance. . . . Nevertheless the occasions, even in research work, in which quantitative data are collected solely with the object of proving or disproving a given hypothesis are relatively rare. Usually quantitative estimates and fiducial limits are required. Tests of significance are preliminary or ancillary.

The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and that is the end of it.

Research workers, therefore, have to accustom themselves to the fact that in many branches of research the really critical experiment is rare, and that it is frequently necessary to combine the results of numbers of experiments dealing with the same issue in order to form a satisfactory picture of the true situation. This is particularly true of agricultural field trials, where in general the effects of the treatments are found to vary with soil and meteorological conditions. In consequence it is absolutely essential to repeat the experiment at different places and in different years if results of any general validity or interest are to be obtained. (Yates, 1951, pp. 32–33)

Yates' closing remarks here make the same appeal, to a method of essentially representative sampling, that was made by Bacon (1627). Gosset was arguing the same point, in his own folksy way, in a letter to Karl Pearson, in the same year that Enlow's book was published:

Obviously the important thing . . . is to have a low real error, not to have a “significant” result at a particular station. The latter seems to me to be nearly valueless in itself. Even when experiments are carried out only at a single station, if they are not mere five finger exercises, they will have to be part of a series in time so as to sample weather and the significance of a single experiment is of little value compared with the significance of the series—which depends on the real error not that calculated for each experiment. But in fact experiments at a single station *are* almost valueless; you can say “In heavy soils like Rabbitsbury potatoes cannot utilise potash manures,” but when you are asked “What *are* heavy soils like Rabbitsbury?” you have to admit—until you have tried elsewhere—that what you mean is “At Rabbitsbury etc.” And that, according to *X* may mean only “In the old cow field at Rabbitsbury.” What you really want to find out is “In what soil and under what conditions of weather do potatoes utilise the addition of potash manures?”

To do that you must try it out at a representative sample of the farms of the country and correlate with the characters of the soil and weather. It may be that you have an easy problem, like our barleys which come out in much the same order wherever—in reason—you grow them or like Crowther's cotton which benefitted very appreciably from nitro-chalk in seven stations out of eight, but even then what you really want is a low real error. You want to be able to say not only “We have significant evidence that if farmers in general do this they will make money by it,” but also “we have found it so in nineteen cases out of twenty and we are finding out why it doesn't work in the twentieth.” To do that you have to be as sure as possible which *is* the 20th—your real error must be small. (quoted in E. S. Pearson, 1939, pp. 247–248)

Even Fisher himself had occasion to lament the ritualistic use to which the newly popular instrument was put.

“High correlations,” and significant results, when obtained, were displayed with some pride, as in some way implying personal competence. When, on the other hand, methods known to be fully efficient are used, the amount of information which the data are capable of yielding has also been assessed, and it is useless either to commend the statistician if it is much, or to reproach him if it is little. The statistician must be treated less like a conjurer whose business it is to exceed expectation, than as a chemist who undertakes to assay how much of value the material submitted to him contains. (Fisher, 1933, p. 46)

Fisher's remarks here about the level of significance being taken as a reflection on the experimenter are interestingly prophetic of psychologists' interpretations

decades later, when many students, and not a few of their advisors, are inclined to take nonsignificant results as an indication of failed research.

7.2.3 *Fiducial Probability*

A friend of mine once gave an undergraduate math exam on which one of the problems was to show that a certain quantity was equal to 2. One hapless student came up with an answer of 3—which, he added hopefully, was “equal to 2 for small values of 3.” Frequentist statisticians have faced the same embarrassing situation, of desperately needing a constant to take on variable values. The unholy grail of statistical inference is to ascribe probability distributions to population parameters. Of all the devices of statistical inference ever invented, fiducial probability promises that objective most directly. There are so many problems with the concept, both logical and technical, that it has never won widespread acceptance; the allure, persisting for nearly a century, was just that it looked so simple. Borrowing an epithet from Mayo (2018), we might say that fiducial probability was Fisher’s most “frequentstein” invention.

Fisher nowhere really gave a definition of fiducial probability, nor a good systematic presentation of the concept or method. As was characteristic of much of his work, he developed the idea mainly through simple examples. When the apparent logic of these examples failed to generalize to more complex problems, Fisher was the only one who could offer any guidance. His personal prestige was enough to keep the theory alive as long as he was, but no longer. The 1963 Conference of the International Statistical Institute in Ottawa included a session on fiducial probability, in which Fisher was to have participated (he died in 1962, while the conference was being planned). Pitman (1963, p. 932) spoke for several of those present in expressing his impatience with the concept and described himself as “uninterested.” And Lindley (1963) moved recognition of what is almost historical fact: “If this meeting were more formal than it is I would propose the motion that this meeting marks the end of the fiducial argument” (p. 920). It is not clear whether Lindley intended the indicative in place of the subjunctive, but he might as well have. The Ottawa Conference was the last major symposium on the concept for over a decade, and the more recent efforts are distinctly post-Fisherian.

Despite its absence from the repertory of contemporary practitioners, fiducial probability is worth discussing both as a distinct approach to statistical inference and as a step in the evolution of the modern theory. Virtually of necessity, the exposition here will follow the model of Fisher’s own. One of his favorite examples for presenting the concept is useful for its simplicity.

Let μ be the median of a distribution of which nothing is known save that its probability integral is continuous (e.g. it need not be differentiable). Let x_1 and x_2 be two observations of the variate; then for any given value of μ it will be true that

- (1) in one case out of 4 both x_1 and x_2 will exceed the median,

- (2) in two cases out of 4, one value will exceed and the other be less than the median,
- (3) in one case out of 4, both will be less than the median.

If a stands for the number of observations less than the median, then a will be a pivotal quantity involving both the unknown parameter and the observations, and having a sampling distribution independent of the parameter, i.e. a takes the values 0, 1, and 2 with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ respectively.

Recognising this property we may argue from two given observations, now regarded as fixed parameters, that the probability is $\frac{1}{4}$ that μ is less than both x_1 and x_2 , that the probability is $\frac{1}{2}$ that μ lies between x_1 and x_2 , and that the probability is $\frac{1}{4}$ that μ exceeds both x_1 and x_2 . The argument thus leads to a frequency distribution of μ , now regarded as a random variate. (Fisher, 1945, p. 131)

Fisher makes explicit here the essential device of fiducial probability:

a process of reasoning by which we may pass, without arbitrariness or ambiguity, from forms of statement in which observations are regarded as random variables, having distribution functions involving certain fixed but unknown parameters, to forms of statement in which the observations constitute fixed data, and frequency distributions are found for the unknown parameters regarded as random variables. (1945, p. 130)

For a more complicated, but more familiar, example, recall Gosset's illustration of the t test with Cushny and Peebles' (1905) data. His presentation has sometimes been described, in fact, as the first example of fiducial inference. In a one-sample t test of whether the sample mean M is significantly above the hypothesized population mean $\mu = 0$, we calculate $t = (M - \mu)/(s\sqrt{n})$, which Gosset found to be equal to $(.75 - 0)/(1.70\sqrt{10}) = 1.40$, corresponding to a p value of .11. Gosset reasoned that, if, on the null hypothesis that $\mu = 0$, the sample mean M has only an 11% probability of lying above .75, ipso facto μ has only an 11% probability of being negative, i.e., more than .75 below M . In such a manner is induced, point by point, a distribution, a *fiducial* distribution, of μ . The pivotal variable in this case is the t statistic, and the fiducial distribution of $(M - \mu)\sqrt{n}/s$ is just the t distribution with $n - 1$ degrees of freedom, or $\mu = .75 + 1.70t_{n-1}/\sqrt{10}$. Essentially what is happening is that, if M and s^2 are the mean and unbiased variance estimate, respectively, of a sample from a normal population, then, as is well known, the quantity $t = (M - \mu)\sqrt{n}/s$ has the Student distribution, and we can easily say with what probability t will exceed a specified value t_0 . But the inequality $t > t_0$ is equivalent to the inequality $\mu < M - st_0/\sqrt{n}$, and so apparently we are making in the same breath a probability statement about the population parameter μ .

Let us take a closer look at the reasoning involved. If we say $p(X < 3) = .2$, then we are obviously entitled to say $p(3 > X) = .2$ —provided we bear in mind that the random variable is X and that we are not making a probability statement about values of 3. That is exactly what confidence intervals do, and why they are so persistently misinterpreted (see below). In fiducial probability, we would like to invert the statement “If $\mu = 0$, then $p(M > .75) = .11$ ” to “If $M = .75$, then $p(\mu < 0) = .11$.“ But from “If A then P” it doesn't follow that “If P then A.” We can set aside the mathematics and look at the problem purely linguistically. Medin and Thau (1992) have provided a perfect example: Compare the sentences “A butcher is like a surgeon” and “A surgeon is like a butcher.” In inverting protasis and apodosis, we need to

constrain the transfer of attributes from butcher to surgeon, or from statistic to parameter. The conceptual problem in the latter case is the attribution of a probability distribution to an unknown constant (facilitated by writing μ or θ in place of a numeral). The main technical problem is that the inversion often yields one-to-many transformations, especially in multiparameter problems; there is no unique pivotal quantity or induced fiducial probability distribution.

As unfamiliar as the concept of fiducial probability is, this line of reasoning may still seem remarkably unremarkable. Fisher (1930) himself suggested that the reason it had previously been overlooked was the focus by Laplace and those following him on the binomial, and he came eventually to the view that the fiducial argument could not be applied to discrete distributions. But one can imagine psychologists encountering fiducial probability here for the first time wondering what all the fuss was about, indeed how such a reasonable sounding procedure could possibly have failed in the history of statistical inference. The observations are after all, they might say, what is given, and real, while the population, being vague and hypothetical, might well be the proper object of a probability distribution.

Whatever plausibility such reasoning holds is testimony to the psychological reality of the epistemic aspect of probability and to the failure of psychological statistics instruction. Without some such device as Bayes' billiard table, population parameters are not the outcome of a random process, and can be described probabilistically only in a purely epistemic sense. But the everyday meaningfulness of this epistemic interpretation has aided the misinterpretation of confidence intervals in a way that apes fiducial probability; hence the deceptive familiarity of the fiducial concept.

Because of the omnipresent risk of interpreting fiducial probability in a Bayesian way, Fisher was at pains in all his writings on fiducial probability to distinguish it from inverse probability. In the case of the t distribution, he insisted:

The distribution which we have obtained is independent of all prior knowledge of the distribution of μ , and is true of the aggregate of all samples without selection. It involves $[M]$ and s as parameters, but does not apply to any special selection of these quantities. To distinguish it from any of the inverse probability distributions derivable from the same data it has been termed the *fiducial* probability distribution, and the probability statements which it embraces are termed statements of fiducial probability. To attempt to define a prior distribution of μ which shall make the inverse statements coincide numerically with the fiducial statements is really to slur over this distinction between the meaning of statements of these two kinds. (Fisher, 1935b, p. 392)

Technically correct though he may be, Fisher is just a little disingenuous in his claim here. It is true that fiducial probability statements apply to the aggregate of all samples, with M and s as parameters for the Student distribution for μ , but in any application we will be inserting numerical values for M and s —with the unsettling implication that the fiducial probability distribution of the population mean, having M and s as parameters, will obviously change from sample to sample. Fisher managed to count this as a point of superiority over the Bayes-Laplace theory of inverse probability:

Whereas . . . the fiducial values are expected to be different in every case, and our probability statements are relative to such variability, the inverse probability statement is absolute in form and really means something different for each different sample, unless the observed statistic actually happens to be exactly the same. (Fisher, 1930, p. 535)

If logically the notion can be accepted of a “frequency distribution” of a population parameter θ changing from one sample to the next, mathematically what is required is a sample quantity, which Fisher called the pivotal quantity, whose distribution is independent of the value of θ but which nevertheless yields information about θ . As he noted in the median example above, a is pivotal: It takes the values 0, 1, 2 with probabilities $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$, respectively, independently of the value of μ ; yet, once the sample is observed, a provides a probability distribution for μ .

A major stumbling block for fiducial probability, however, is that different pivotal functions yield different fiducial distributions, and some principle or set of criteria is needed to select the “correct” pivotal. Fisher’s answers—especially the stipulation that only “natural” pivots could be used—tended to be vague and unsatisfactorily ad hoc. The most important requirement he ultimately specified was the condition of sufficiency, a concept he had introduced in his “Mathematical Foundations” paper (1922b).

If T is a sufficient statistic and T' any other statistic whatever, then the simultaneous distribution of T and T' in random samples shall be such that for any chosen value of T the distribution of T' shall be wholly independent of the value of the parameter θ to be estimated. In other words, as soon as we know the value of T , then the value of T' is completely irrelevant to the estimation of θ , and supplies no information whatever about it. Since, in the case of sufficient statistics, this is true of all alternatives which can be proposed, we are using our words consistently when it is said that the sufficient statistic contains all the information about θ originally present in the data. (Fisher, 1951, pp. 47–48)

For a normal distribution, the mean and variance are jointly sufficient statistics, since they suffice to specify a normal distribution exactly. Fisher’s argument here was that:

It is essential to introduce the absence of knowledge a priori as a distinctive datum in order to demonstrate completely the applicability of the fiducial method of reasoning to the particular real and experimental cases for which it was developed. (Fisher, 1956/1973, p. 59)

For, although in the deduction of statements of certainty it is legitimate to draw inferences from some of the axioms available while ignoring others, or, in other words to base a valid argument on a chosen subset only of the available axioms, no such liberty can be taken with statements of uncertainty, where it is essential to take the whole of the data into account, though some part of it may be shown on examination to be irrelevant, and not to affect the result. (p. 58)

The restriction to sufficient statistics unfortunately proved very narrow; when the condition of consistency is added, it turns out that fiducial distributions can be obtained only for normal and gamma distributions, which Fraser (1963) called “an extremely limited range of problems” (p. 843).

Further technical problems emerged with attempts at generalization to multivariate problems with more than one parameter. The bivariate normal distribution has

five parameters—mean and variance for X and Y , plus the correlation. Seidenfeld (1979) shows that fiducial inference depends on the order in which these parameters are taken up. Wilkinson (1977) shows that it depends on whether the discriminant function is specified a priori or is optimized from the data.

In addition to the challenging technical problems, the question also remained as to exactly what fiducial probability represented. Fisher's original term was "fiducial probability distribution," and he said once (1936b) that "This distinctive terminology is not intended to suggest that fiducial probability is not in the strictest sense a mathematical probability, like that of any other to which the term ought to be applied" (p. 253). But Kyburg Jr. (1963a) notes that:

The fact that fiducial probability statements cannot be given the kind of empirical content that most statisticians still like to attribute to probability statements has led some writers to insist that there is a difference between fiducial probability and other kinds of probability. Even writers who are decidedly on Fisher's side take this view. (p. 895)

He goes on to quote Mahalanobis' laudatory biographical sketch (1938, p. 270): "Fisher has laid emphasis on the fact that the concept of fiducial probability, *though entirely different from ordinary probability, is equally rigorous*" (Kyburg's emphasis). Fisher, in an attempt to dispel some of the confusion, spoke, in his later writings, simply of a "fiducial distribution"; but Kendall (1963) remarks, with a certain edge of desperation, that:

it is not a probability distribution to anyone who rejects Bayes's approach, and indeed, may not be a distribution of anything. Fisher nevertheless manipulated it as if it were, and thereafter maintained an attitude of rather contemptuous surprise towards anyone who was perverse enough to fail in understanding his argument. (p. 4)

Leonard Savage (1961), who was as responsible as anyone else for introducing Bayesian statistics to American psychologists, described fiducial probability as "a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs" (p. 568). On the question of what the concept of fiducial probability is supposed to do, he said:

I asked Fisher once and he gave the very candid answer: "I don't understand yet what fiducial probability does. We shall have to live with it a long time before we know what it's doing for us. But it should not be ignored just because we don't yet have a clear interpretation." I respected that answer, though I am coming to have a certain impatience with it. (1963, p. 926)

The concept of fiducial probability is not a familiar one to contemporary social scientists; it was known only to about one generation of students, approximately in the 1940s. Professional statisticians have largely ignored it, too. Frequentists disdained it as unintelligible and incoherent, and Bayesians had no need of it. As late as 2014, Oliver Mayo described fiducial probability as "dead." Yet, despite—or perhaps because of—the deep ambiguities, both logical and mathematical, in the concept of fiducial probability, it is so close to the heart of statistical inference, especially from other than a Bayesian point of view, that the temptation to rehabilitate the concept has remained strong. In the years since Fisher's death, Hacking (1965), Fraser (1968), Kyburg Jr. (1974), Wilkinson (1977), Seidenfeld (1979), and Hannig

(2009) have made notable attempts at systematizing the fragments Fisher left behind. Fraser's program of "structural inference," based on groups of functions invariant under transformation, remains at an abstract and mathematically formidable level; A. W. F. Edwards (1977), frustrated with the mathematical difficulty of structural probability, referred to the theory as a "black hole," because theorists who got into it seem to have forgotten the route by which they had traveled. Seidenfeld placed it beyond the scope of his study, but he criticized it on the same grounds as Hacking's and Kyburg's theories, for failing to provide a means for the unique representation of ignorance. Seidenfeld's fatal example is similar to the Amtrak example used in Chap. 6 to illustrate the difficulties of uniform prior probability distributions in the classical Bayes-Laplace approach; he showed that estimates of the volume of a hollow cubic structure variously made from weighing either the liquid filling it or the weight of a rod laid along its edge and cut to the same length are both valid and incompatible. Seidenfeld didn't offer a solution himself.

Fraser (1968) got around the problem of the nonuniqueness of pivots, and corresponding fiducial distributions, by making the pivotal variable part of the model specification. Wilkinson (1977), for his part, took the bull by the horns and rejected the fundamental notion that the evidential implications of a body of data ought to be unique. His version of fiducial probability is based on his "Noncoherence Principle: The inferential implications of observational data alone (with sampling distribution known in parametric form) are noncoherent, in that they cannot be represented by a single inferential probability distribution on the parameter space" (Wilkinson, 1977, p. 121). He dealt with the problem of nonuniqueness of fiducial solutions, in other words, by embracing it as a necessary and realistic feature. Just as, perhaps, we have to specify measurement scales and bring empirical knowledge to bear in application of the principle of indifference, so use of fiducial probability may require that we specify the appropriate questions and procedures.

A fundamental property of "ignorance" is its preservation under transformation—ignorance about θ implies ignorance about any function of θ , and one would expect the same to hold approximately for "near-ignorance." Such a property is compatible with the view that belief is intrinsically noncoherent, but not with the Bayesian view that belief is representable by a single probability distribution. I believe, indeed, that the Noncoherence Principle is as fundamental in Statistical Inference as the Heisenberg Uncertainty Principle is in Quantum theory—the parallel is unusually apt. (p. 142)

What is especially interesting about Wilkinson's theory is that his concept of non-coherence rests, perhaps a little more than he acknowledged, on Shafer's (1976) theory of nonadditive belief functions (see Chap. 10). One interesting consequence is that fiducial probability, in Wilkinson's construction, does not obey the Kolmogorov axioms⁵ (pp. 121, 166). Hannig (2009), while acknowledging the

⁵According to A. W. F. Edwards (1977), Fisher was once asked, following a lecture, whether fiducial probability followed the Kolmogorov axioms. "The reply—following a long pause—was 'Kol who?'" (p. 145). Kolmogorov's (1933/1950) axiomatization of probability was greeted very eagerly by twentieth-century empiricists. But it covered only probabilities that could be mapped onto relative frequencies; the requirement that probabilities of a proposition and its denial had to

interest of Wilkinson's solution, prefers, like a Bayesian, to leave no probability unallocated. He has done much work to try to systematize fiducial probability, developing rules to ensure uniqueness (though not being able to do anything about the conceptual problem). It is tempting to speculate on how different the history of statistical inference—and research in the social sciences—might have been had Hannig's work been accomplished in the 1930s. But I suspect the answer is: Not much. Fiducial probability offers at best what confidence intervals are taken to be, and, as such, would have remained distantly second to significance testing, which appeared to offer an unambiguous answer to the question of whether a given result was important.

7.3 Neyman and Pearson

Whereas Fisher was active all his life not only in mathematical statistics but in experimental and theoretical scientific work and to some extent in philosophy of science, Neyman was more the pure mathematician. He went into agricultural statistics in Poland primarily for financial support. In 1925 the Polish government sent him to work for a year in Karl Pearson's laboratory at University College, London. He had been excited as a student by the positivism of Pearson's *The Grammar of Science* (1892), but very quickly encountered the master's intolerance of criticism. A paper Neyman had published in Poland pointed out a mistake in one of Pearson's papers; when Pearson read it, he declared that Neyman was wrong; when Neyman dared to maintain, in front of other students, that he was right, Pearson thundered: "That may be true in Poland, Mr. Neyman, but it is not true here!" (Reid, 1982, p. 57). Partly through Egon's intervention, Pearson was ultimately brought around to Neyman's point of view, and Neyman was allowed to stay.

Interestingly, Neyman and Fisher, 4 years his senior, got along well for the first few years, after Gosset introduced them in 1926. Neyman made some efforts to bring Fisher together with Egon, who was intimidated by him (Reid, 1982).⁶ When Fisher (1935a) presented his paper on "The Logic of Inductive Inference" to the Royal Statistical Society, Neyman was the only discussant who wasn't hostile, and Fisher was grateful. Neyman soon found, however, that Fisher was no more tolerant of criticism than Karl Pearson. In a paper to the Statistical Society in 1935, Neyman (Neyman, Iwaszkiewicz, & Kolodziejczyk, 1935) criticized Fisher's test for Latin squares. According to Oscar Kempthorne, a distinguished student of Fisher's,

sum to 1 thus excluded nonadditive probabilities in particular (see Chap. 10). But the fact that it covered both frequentist and Bayesian probabilities allowed proponents of both to say that they accepted the same axiom system and differed only in its interpretation.

⁶Egon had incurred Fisher's wrath in publishing a review of *Statistical Methods for Research Workers in Nature*. Gosset thought the review fair, but Fisher was offended and wrote an angry response. Gosset wrote a conciliatory letter himself, but, constrained to anonymity by Guinness, used the Galton Laboratory as his address, which incensed Karl Pearson (Tankard, 1984).

Neyman had the better point (Reid, 1982), but Fisher opened the discussion with a sarcastic remark that Neyman should have chosen a topic on which he could speak with authority. One evening not long afterward,

Neyman and Pearson returned to their department after dinner to do some work. Entering, they were startled to find strewn on the floor the wooden models which Neyman had used to illustrate his talk on the relative advantages of Randomized Blocks and Latin Squares. These were regularly kept in a cupboard in the laboratory. Both Neyman and Pearson have always believed that the models were removed by Fisher in a fit of anger. (Reid, 1982, p. 124)

A few weeks later Neyman applied for an advertised readership in statistics at University College. Egon Pearson, when he took over the Department of Applied Statistics in 1933, had managed to squeeze one member out of the department to make room for Neyman, thus ending the latter's chronically precarious professional and financial situation in Poland; but the readership held the advantage of being a tenured position. Fisher's response, following Egon's model, was to advise Neyman that he should lecture only from Fisher's *Statistical Methods for Research Workers*. When Neyman refused, Fisher promised to oppose him in every way he could (Reid, 1982).

When Neyman finished, in 1936, a major paper on estimation, he naturally assumed that Egon, having taken over the editorship of *Biometrika* after his father's death, would publish it in that journal; but Egon, curiously, rejected it as too long and mathematical. With *Biometrika* closed to him by his good friend and Fisher, he thought, controlling everything else, Neyman felt close to despair. He devised, however, a shrewd strategy which succeeded: He sent it to Harold Jeffreys, the leading Bayesian, for submission to the Royal Society. Jeffreys was later to describe his own position as closer to Fisher's than to Neyman's, but Neyman guessed correctly that Jeffreys was enough of an enemy of Fisher's to sponsor his paper. To Neyman's surprise, Fisher was not selected as a referee, and the paper was accepted for publication in the *Philosophical Transactions* (Neyman, 1937a).

While that paper was under review, Neyman made a successful tour of the United States, shortly following the second made by Fisher. Neyman's personal manner served him better than Fisher's. Fisher would very likely have accepted an offer from Berkeley but had thoroughly alienated people there with his insensitivity and conceit (Reid, 1982). Neyman, however, was offered and accepted a professorship at the University of California in 1938. He was also courted by the University of Michigan, which was the leading center of academic statistics at the time, but he chose Berkeley specifically because there was nothing else there in statistics (Reid, 1982). All the same, he felt a great reluctance to come to the United States at all, because he was aware of how far American students were below the academic level of Europeans. Statistics courses in most departments at Berkeley required only high school algebra; and it is not hard to imagine Neyman's reaction on finding Griffith Evans, chair of the mathematics department, sorting students into sections according to whether they could add fractions. Apart from his being hemmed in professionally in England, the deteriorating situation in Poland in 1938 was a deciding factor in his move.

At least three ironies marked Neyman's tenure at Berkeley. One was that in 1942 Alfred Tarski, who had been working on his Ph.D. at the university in Warsaw at the same time as Neyman, joined the mathematics department at Berkeley; yet their views were so opposed that their colleagues referred to them as "Poles apart." Another was the object of a curious agreement with Fisher: The world's two leading statisticians, both heavy smokers, disputed all their lives the link between smoking and cancer.⁷ The third is sadder, showing Neyman following the tradition of Fisher and Karl Pearson. Erich Lehmann, perhaps Neyman's most distinguished student, eventually published the definitive work on the Neyman-Pearson theory (1959). Because it extended his work slightly, however, Neyman would not allow Lehmann to teach unless he could review the manuscript. So Lehmann did not teach again until Neyman was no longer chair (Reid, 1982).

7.3.1 Hypothesis Testing

The active collaboration between Neyman and Pearson lasted only a few years, and much of it took place through correspondence, when Neyman was in Poland or France. It evidently began with the young Pearson's dissatisfaction with the currently ad hoc state of statistical science, when tests were chosen on a largely intuitive basis, and especially with his doubts about the small-sample theory of Fisher and Gosset. He wrote his concerns to Gosset, who supported his efforts and also mentioned the relevance of alternative hypotheses in the definition of a best test. There are some situations, Gosset pointed out, where extremely improbable results would not occasion rejection of the null, just because the alternative was even more incredible. If, for example, he had shuffled the cards before dealing himself a hand of 13 trumps, even the astronomical odds against that event could not persuade him that it was other than chance (E. S. Pearson, 1939).

Gosset, no more a mathematician than Pearson, was not the man to help him. The appeal of Neyman, whom he was at this point just getting to know, may have had something to do with the fact that Neyman was specifically an outsider, as yet unembroiled in the controversies with either Fisher or the older Pearson.

⁷The controversy persisted as long as it did because of the lack of experimental evidence: No one could produce cancer in laboratory animals by cigarette smoke. That may have been just because linoleic acid, a polyunsaturated fatty acid present in vegetable oils, is necessary in their diet before they can get cancer (Tucker Goodrich - How Linoleic Acid Wrecks Your Health (mercola.com)). The epidemiological link was publicized by Richard Doll (Doll & Hill, 1950), but the data had actually been collected in the 1920s. The man who spearheaded that project was not a doctor, but a well-known health advocate: vegetarian, teetotaler, and of course nonsmoker—who strongly suspected that cigarette smoking was harmful and directed the doctors under his control to collect the relevant data: Adolf Hitler. When they compiled the data, the answer fell out glaringly. But after the war, no one wanted to look at Nazi research (except for rockets), if they could even read it—except for Richard Doll, who used it without attribution (Kealey, 2021).

In a retrospective account of the beginning of their collaboration, Pearson (1962) listed several specific historical influences on the shape of their theory, ideas which were current in the middle 1920s:

- (a) The way of thinking which had found acceptance for a number of years among practising statisticians, which included the use of tail areas of the distributions of test statistics.
- (b) The classical tradition that, somehow, prior probabilities should be introduced numerically into a solution—a tradition which can certainly be traced in the writings of Karl Pearson and of Student, but to which perhaps only lip service was then being paid.
- (c) The tremendous impact of R. A. Fisher. His criticism of Bayes's Theorem and his use of Likelihood.
- (d) His geometrical representation in multiple space, out of which readily came the concept of alternative critical regions in a sample space.
- (e) His tables of 5 and 1% significance levels, which lent themselves to the idea of choice, in advance of experiment, of the risk of the “first kind of error” which the experimenter was prepared to take.
- (f) His emphasis on the importance of planning an experiment, which led naturally to the examination of the power function, both in choosing the size of sample so as to enable worthwhile results to be achieved, and in determining the most appropriate test.
- (g) Then, too, there were a number of common-sense contributions from that great practising statistician, Student, some in correspondence, some in personal discussion. (p. 395)

These factors provided some initial influences and directions, but the final conception of the theory came about gradually. A chronological survey of the writings of Neyman and Pearson conveys a strong impression of openness to many alternative routes in the beginning stages, but their evolving commitment to a particular line of approach led them, as they followed through its implications, to a progressive radicalization of the theory. These theoretical developments were primarily the contribution of Neyman.

The progression can be seen clearly in their attitude toward inverse probability. Both Neyman and Pearson authored papers on the approach early in their careers. Pearson (1925) did an empirical study of Bayes' Theorem, in which he argued for provisional acceptance of the theorem as a basis for statistical inference, where there was some positive evidence for the prior distribution assumed; and Neyman (1930) showed that the Bayesian approach was feasible in large samples, when the exact form of the prior distribution became unimportant, and applied the method to a number of standard estimation problems. Whatever opposition they had to inverse probability had a more practical sound to it: The problem that prior probabilities seldom existed in frequency form. Later their opposition acquired a more philosophical character, as their theory became tied more explicitly to a frequency theory of probability.

To American psychologists and other users of the theory brought up on the finished doctrine, the broad-mindedness of their early papers would surely be breathtaking. Despite its having been accepted as virtually *the* theory of hypothesis testing, its authors did not put it forth in those terms. Indeed, in their first joint paper (Neyman & Pearson, 1928a), they presented their approach, based on integrals of probability contours, merely as one of four or five possible alternatives, such as inverse probability and the method of maximum likelihood. With characteristic modesty and circumspection, they placed all the methods on more or less an equal

footing: “If [these methods are] properly interpreted we should not describe one as more *accurate* than another, but according to the problem in hand should recommend this one or that as providing information which is more *relevant* to the purpose” (Neyman & Pearson, 1928a, pp. 230–231). Moreover,

The process of reasoning . . . is necessarily an individual matter, and we do not claim that the method which has been most helpful to ourselves will be of greatest assistance to others. It would seem to be a case where each individual must reason out for himself his own philosophy. (p. 230)

As this statement clearly indicates, they were at that time still placing their theory in the traditional domain of inference and reasoning; but it was in this area that their theory was to undergo the most striking transformation. Thus they went on, in Part II of their first joint paper, to characterize the object of their theory as “problems of inference” (Neyman & Pearson, 1928b, p. 263), and Pearson (1962) describes their intentions in the following unmistakable terms: “We were seeking how to bring probability theory into gear with the way we think as human beings” (p. 395). Interestingly, he gives here no indication, save possibly the past progressive tense, that their program ever changed in this respect.

It is true that they never thought of their theory as exhaustive of the process of reasoning, or even of decision.

We were certainly aware that inferences must make use of prior information and that decisions must take account of utilities, but after some considerable thought and discussion round these matters we came to the conclusion, rightly or wrongly, that it was so rarely possible to give sure numerical values to these entities, that our line of approach must proceed otherwise. Thus we came down on the side of using only probability measures which could be related to relative frequency. Of necessity, as it seemed to us, we left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters—to use our terminology—as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities. (Pearson, 1962, pp. 395–396)

Their first joint paper actually made it quite clear at the time that an important part of the process of inference or decision was left out of the formal theory:

In many cases there is probably no single “best” method of solution. The sum total of the reasons which will weigh with the investigator in accepting or rejecting the hypothesis can very rarely be expressed in numerical terms. All that is possible for him is to balance the results of a mathematical summary, formed upon certain assumptions, against other less precise impressions based upon *a priori* or *a posteriori* considerations. The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision; one man may prefer to use one method, a second another, and yet in the long run there may be little to choose between the value of their conclusions. What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitations, should be clearly understood. (Neyman & Pearson, 1928a, p. 176)

As Pearson’s statement in the preceding paragraph makes clear, it was their commitment to the frequency conception of probability which dictated that certain aspects of the inference process be left out of the theory of hypothesis testing—namely, those aspects which were not (at least not obviously) subject to numerical representation. In time they ceased to refer to inference altogether, as their theory

assumed a behavioristic orientation. Neyman (1938) coined the term “inductive behavior” as the new object of the theory and declared that “the theory of testing hypotheses has no claim of any contribution to the [theory of] inductive reasoning” (Neyman, 1942, p. 301). Hypotheses would be accepted or rejected, ‘But to decide to ‘affirm’ does not mean to ‘know’ or even to ‘believe’’ (Neyman, 1938, p. 56).

The change really came about when Neyman set out to ground their theory formally in the frequency theory of probability. His ideas were presented most fully in his paper “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability” (1937a). Neyman’s aligning himself with the classical theory is a little curious; it indicates that he is taking *classical* to refer to the frequency theory in the broadest sense: The classical theory defined probability as the ratio of favorable to possible cases, and was thus limited to discrete probabilities, whereas most of Neyman’s applications were to continuous distributions. Neyman was one of the translators of von Mises’ *Wahrscheinlichkeit, Statistik und Wahrheit*, which, he says, “confirmed me as a radical ‘frequentist,’ intent on treating probability as a mathematical idealization of relative frequency. However, von Mises’ definition of probability did not attract me and I became a follower of Kolmogorov” (Neyman, 1967a, p. ix).

The Concept of Probability Neyman’s writings overall, like Fisher’s, evidence an extraordinary struggle with the concept of probability. Whereas Fisher claimed to be a frequentist, but in application resorted to an epistemic or subjective interpretation of probability, Neyman was a thoroughgoing frequentist in application, but, when he was writing about the concept itself in the abstract, tried to present a sheerly axiomatic approach. To compound the confusion, Neyman’s illustrative examples tended to be simple Laplacean probabilities based on the principle of indifference. Both men tended to be more sensitive to the inconsistencies in the other’s position than to their own.

In his formal discussion of probability, Neyman, following the axiomatic tradition inaugurated by Kolmogorov (1933/1950), introduced the idea of a “fundamental probability set” (e.g., the real line) and some measure over it which satisfies the Kolmogorov axioms. For an infinite F.P.S. there are infinitely many possible measures, the choice among which is mathematically arbitrary; Neyman chose the Lebesgue integral (which is sophisticated, and allows for freakish cases, but which is equivalent to the ordinary Riemann integral in nearly all applied problems). If A is the F.P.S. and B is some subset, then the probability of B is defined, in a predictable way, as the ratio of the measure of B to the measure of A. Neyman went on briefly to consider Bertrand’s paradox of the chord chosen at random in a circle (Chap. 6), and reached what was Keynes’ (1921/1973) solution, that the various solutions represent answers to different questions. But the problem also gave him the occasion to emphasize that:

The conception of “equally probable” is not in any way involved in the definition of probability adopted here, and it is a pure convention that the statement “In picking up at random a chord, we first select a direction of radius, all of them being equally probable and then we choose a distance between the centre of the circle and the chord, all values of the distance

between zero and r being equally probable" means no more and no less than "For the purpose of calculating the probabilities concerning chords in a circle, the measure of any set (A_1) of chords is defined as that of the set (A'_1) of points with coordinates x and y such that for any chord A_1 in (A_1), x is the direction of the radius perpendicular to A_1 and y the distance of A_1 from the centre of the circle. (A_1) is measurable only if (A'_1) is so." (Neyman, 1937a, pp. 338–339)

It is perhaps his commitment to the axiomatic approach to probability which allowed him to appear grandly indifferent to empirical interpretations of probability in one statement:

It may be useful to point out that although we are frequently witnessing controversies in which authors try to defend one or another system of the theory of probability as the only legitimate, I am of the opinion that several such theories may be and actually are legitimate, in spite of their occasionally contradicting one another. Each of these theories is based on some system of postulates, and so long as the postulates forming one particular system do not contradict each other and are sufficient to construct a theory, this is as legitimate as any other. In this, of course, the theories of probability are not in any sort exceptional.

Both Euclidean and non-Euclidean geometries are equally legitimate, but, e.g. the statement "the sum of angles in a linear triangle is always equal to π " is correct only in the former. In theoretical work the choice between several equally legitimate theories is a matter of personal taste only. In problems of application the personal taste is again the decisive moment, but it is certainly influenced by considerations of the relative convenience and the empirical facts. (Neyman, 1937a, p. 336n)

But up to this point, Neyman's essay remains purely an exercise in assimilation.

A theory of statistics for application, however, cannot be merely stipulative; it requires accommodation to real events. The real events he has in mind are "random experiments," such as the drawing of a random sample from a population. The concept of randomness, it should be noted (though Neyman does not), is intimately connected, by his own definition, with the concept of equiprobability, from which he wanted to dissociate himself: In the case of an infinite population, "When we speak of a random sample we mean that it is drawn so that (1) the probability of each individual of the population being included in the sample is the same, (2) separate drawings are mutually independent" (Neyman, 1937a, p. 335). Apparently he intends for his definition of probability to be independent of any Laplacean assumption of equiprobable alternatives and to represent merely an arbitrary measure on some set; the concept of equiprobability will simply be built into the choice of some fundamental probability sets and their measures, and will not require separate designation. But the only objects he applies his theory to are random experiments, and randomness cannot evidently be defined without reference to equiprobability. If randomness is to be understood in its usual (somewhat inscrutable, but not arbitrary) meaning, then so must the concept of equiprobability and of probability more generally. At the point of application, all the beautiful arbitrariness and avoidance of traditionally troublesome assumptions vanish, and we are squarely, if unadmittedly, back in familiar territory.

Let us hear Neyman's account of the transition.

The justification of the way of speaking about the definition of the measure within the fundamental probability set in terms of imaginary random experiments lies in the empirical

fact, which Bortkiewicz insisted on calling the law of big numbers. [The Polish nationalism is pardonable, but Poisson still beat him by 80 years.] This is that, given a purely mathematical definition of a probability set including the appropriate measure, we are able to construct a real experiment, possible to carry out in any laboratory, with a certain range of possible results and such that if it is repeated many times, the relative frequencies of these results and their different combinations in small series approach closely the values of probabilities as calculated from the definition of the fundamental probability set. (Neyman, 1937a, p. 339)

He goes on to acknowledge a certain fiction in the “empirical fact”:

e.g., if we take any coin and toss it many times, it is very probable that the frequency of heads will not approach $\frac{1}{2}$. To get this result, we must select what could be called a well-balanced coin and we have to work out an appropriate method of tossing. Whenever we succeed in arranging the technique of a random experiment, say E , such that the relative frequencies of its different results in long series sufficiently approach, in our opinion, the probabilities calculated from a fundamental probability set (A), we shall say that the set (A) adequately represents the method of carrying out the experiment E . The theory developed below is entirely independent of whether the law of big numbers holds good or not. But the applications of the theory do depend on the assumption that it is valid.⁸ (pp. 339–340)

Thus the theory applies only to “well-balanced” coins, and the problem is then to say how to specify in advance that a coin is well-balanced without recourse to equi-probability principles; as we shall see, specification of probability models in advance is a key feature of the Neyman-Pearson theory. It is difficult to believe that Neyman was not appealing to the principle of indifference in the following example from his *First Course in Probability and Statistics*: “An ordinary die has six sides. Hence the F.P.S. is composed of $n(A) = 6$ elements. Only one of the sides has six dots on it. Hence $n(AB) = 1$. Hence $P_1 = 1/6$ ” (Neyman, 1950, pp. 16–17). Taking this example out of context, which is often done in criticisms of Neyman, is somewhat unfair, however, for he goes on to contrast this probability with the probability of the same die falling with six dots up. The F.P.S. for the latter example is evidently a hypothetical set of all possible tosses of this die. A problem with Neyman’s treatment at this point is that, while he defines probability by reference to the F.P.S., he nowhere defines that concept. It appears from his discussion that a fundamental probability set may be any set whatever, and the probability associated with any of its members is simply the proportional representation of that kind of element in the set.⁹

⁸The reference to “our opinion” as the criterion of convergence explicitly acknowledges the subjectivity of the frequency definition that was discussed in Chap. 6.

⁹It is relevant in this context to mention an interesting criticism of the classical definition of probability by Jeffreys (1939/1961).

It often gives a definite value to a probability; the trouble is that the value is one that its user immediately rejects. Thus suppose that we are considering two boxes, one containing one white and one black ball, and the other one white and two black. A box is to be selected at random and then a ball at random from that box. What is the probability that the ball will be white? There are five balls, two of which are white. Therefore, according to the definition, the probability is $2/5$. But most statistical writers, including, I think, most of those that professedly accept the definition, would give $(\frac{1}{2})(\frac{1}{2}) + (\frac{1}{2})(\frac{2}{3}) = 5/12$. This follows at once on the present theory, the terms representing two applications of the product rule to give the probability

To recapitulate, Neyman defines probability as a measure on a certain set; he then asserts that, by Bernoulli's Theorem, random experiments will conform to this definition in their long-run behavior. At this point the definition is frozen into everyday, rather than arbitrary mathematical, terms, and he faces the question of how to handle discrepancies. In actual practice, of course, if empirical frequencies, even over a hundred trials, depart sufficiently from the hypothesized probability, that hypothesis is rejected in favor of an alternative—that is what significance testing is all about; but, as Jeffreys (1939/1961) insists, nothing in Neyman's theory of probability allows for such a move. The occurrence of any event, so long as it is included somewhere in the F.P.S., is consistent with the model; the model can be impugned only if we expect low-probability events not to occur, and expectation and belief are precisely what the definition excludes. As Hacking (1965) puts it, "The Neyman-Pearson theory defends chance in terms of chance. To stop that spiral, one needs the logic of support" (p. 105). Lucas (1970) makes a similar point: Having defined probability as relative frequency, Neyman must count on empirical frequencies approximating values that are plausible on other, nonfrequency grounds; in the event of discrepancy between theoretical and empirical probabilities, it is the definition which has to give. We can save the definition of probability only by freeing it of its empirical bonds and attaching it to an ideal (a "well-balanced" coin, an *unbiased* die, a *random* sample)—but then we are appealing to something else in place of frequency for our definition of probability. The reductionist argument of Neyman and Pearson, and others like them, is seductive precisely because we, as readers, are tacitly supplying the meaning—in terms of belief, expectation, or support—which they are denying to the concept of probability.

Implications for Statistical Inference Whether or not the frequentist program can logically succeed, the commitment to frequency probabilities carries several implications for the shape of a theory of statistical inference or decision, all of them rather closely related.

of drawing each of the two white balls. These are then added by the addition rule. But the proposition cannot be expressed as the disjunction of five alternatives out of twelve. My attention was called to this point by Miss J. Hosiasson. (p. 370)

Neyman (1952) replied to this criticism by contending that the elements of the F.P.S. in this case are not balls but pairs of random selections, of which 12 may be enumerated. His point is well taken, though it is not clear what in his own definition of probability enables him to argue for this particular F.P.S. as *the* correct one, other than the fact that it affords a convenient match with the classical answer given by the principle of indifference.

There is more to the story, however: Neyman goes on to say that Janina Hosiasson was one of his assistants in Warsaw, "a very talented lady who has written several interesting contributions to the theory of probability" (1952, p. 12), one of them dealing specifically with paradoxes that arise from imprecisely stated conditions. "In these circumstances," he says:

It is most unlikely that Miss Hosiasson could fail in the application of the direct method to a simple problem like the one described by Dr. Jeffreys. On the other hand, I can well imagine Miss Hosiasson making a somewhat mischievous joke. (p. 13)

Miss Hosiasson herself could not resolve the matter; Neyman mentions in the dedication of the book that she was murdered by the Gestapo in World War II.

The most important of these is a strong pull toward a decision theory in place of a theory of inference. That feature was already present in Fisherian significance testing; the rejection of the null hypothesis or the failure to reject it circumvents the problem of assigning probabilities to hypotheses, which is difficult under the frequency theory of probability. Fisher equivocated, however, wanting to retain an epistemic or evidential interpretation of significance levels, and fiducial probability made the hedge explicit.

Neyman and Pearson (1933a) describe, in one of their early papers, how they were led from rules of inference to rules of behavior.

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a “rule of behaviour”: to decide whether a hypothesis, H , of a given type be rejected or not, calculate a specified character, x , of the observed facts; if $x > x_0$ reject H , if $x \leq x_0$ accept H . Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$ or false when $x > x_0$. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. (pp. 290–291)

On the frequency theory, probability statements—significance levels, in particular—must refer to clearly specified classes (in this case, indefinitely extended sequences of tests under the same conditions), and such statements applied to individual tests carry no meaning, except in the usual frequency sense of reference to that infinite set.¹⁰ Neyman (1937a) contended that it is not logically necessary that the reference class of a statement of significance be a set of repetitive tests on the same hypothesis. Rather all that is required is that the experimenter choose a number between 0 and 1—a small number for a level of “significance” or a large number for a “confidence” coefficient—and stick to this as his or her “operating characteristic.” Then, over a very long sequence of tests, even if the parameters and distributions are different, the α will remain the same; and, by Bernoulli’s Theorem, the long-run frequency of false rejections should be approximately α . A further requirement, which Neyman does not state, is that these successive experiments be independent; there appears to be some ambiguity in interpreting this requirement

¹⁰The set designated by the significance level is obviously infinite when the sampling distribution is continuous, and it might be thought that a test involving a discrete distribution (e.g., the sign test) would refer to a finite set. The ambiguity here is the same as that between 1/6 as the proportion of sides of a die bearing a 6 and 1/6 as the long-run relative frequency of tosses turning up a 6. The former is of course the “classical” probability and the latter the frequentist meaning, and Neyman (1937a) has enhanced confusion by referring to his theory as “classical.” But his writing generally makes it clear that probability, for him, involves the proportion of outcomes over an indefinitely long run. Hence, though the tail of a binomial distribution contains but a finite set of points, these are ideally, or logically, defined and represent the proportion of false rejections to which an actual sequence of tests should converge.

for an extended program of research on theoretically related hypotheses. This problem has apparently never been discussed, but the question is largely academic. In any application, the reference set is a portion of a sampling distribution, representing hypothetical tests on the same hypothesis (different independent samples from the same population).

The focus on decisions in repetitive circumstances was in itself sufficient to lead Neyman and Pearson, already by the middle 1930s, to emphasize problems of quality control in their illustrative examples. The situation in manufacturing, in fact, is well suited to such a paradigm: The interest there is exclusively in long-run profit and loss, and not in the correctness or reasonableness of any individual decision along the way.

The focus on “rules of behavior” as the object of the theory entailed a commitment to careful planning of research.

We were regarding the ideal statistical procedure as one in which preliminary planning and subsequent interpretation were closely linked together—formed part of a single whole. It was in this connexion that integrals over regions of the sample space were required. Certainly, we were much less interested in dealing with situations where the data are thrown at the statistician and he is asked to draw a conclusion. (Pearson, 1962, p. 396)

Their choice of the integral of tail probabilities as the rejection criterion, which Jeffreys (1939/1961) criticizes, was to some extent conditioned by current statistical practice, but it was well suited to an approach where conditions were to be laid out in advance. Unlike the likelihood ratio, which could take any value, significance levels easily admitted of standardization for comparisons across tests—indeed Fisher’s tabled .05 and .01 values already offered a convenient suggestion.

Interestingly, this particular aspect of the Neyman-Pearson theory—the commitment to planning, essentially—was to become the basis of one of the major criticisms of the theory, by Hacking (1965). He argues essentially that procedures which are appropriate for “before-trial betting on hypotheses” are not (necessarily) well suited for “after-trial evaluation of results” and that the Neyman-Pearson theory is geared exclusively to the former, whereas the latter may be more appropriate for scientific research. More precisely, the Neyman-Pearson theory is appropriate for after-trial evaluation provided that we do not discriminate between outcomes in the rejection region, which is ordinarily tantamount to discarding data. So long as we know, after the experiment, only that the result fell in the rejection region, and not what the actual outcome was, Neyman-Pearson theory may be regarded as suitable also for after-trial evaluation. This situation may well obtain in manufacturing or actuarial applications, where there is no concern with the individual result, and these are in fact the paradigmatic and most defensible applications of the theory; it is arguably not the situation in scientific research, as will be discussed further in Chap. 9.

The focus on planning of experiments further entails, in particular, that the rejection region be specified in advance. Otherwise the test has no determinable size, and the significance level, at least on a frequency interpretation, loses its meaning. Allowing ourselves to select the rejection region (alpha level, one tail versus two) after the data have been collected is analogous to the luxury of declaring the wild

cards after we see our hand. No doubt some kind of game could be played this way, and there is no shortage of psychologists willing to be partners in this sort of enterprise; but what is not allowed any longer, in logic, are claims about the rarity of the various events we witness. The important difference between poker and significance testing is that statements of probabilities are not the primary object in the former, as they ostensibly are in the latter, where it is more difficult to see just what kind of game is left.

If we care at all about the magnitude of effects, and not merely their significance, then our advance specification must also include the sample size, for it is the sample size, jointly with everything else, which determines the power of the test, the size of an effect to which the test will be sensitive. Hogben (1957) was making the point more than 60 years ago; it is now appearing in more of the standard statistics texts (e.g., Hays, 1963, 1981, 1988).

Implicit in the foregoing is the necessity of specifying alternatives in the construction of a test. When integrals across probability contours are to be used as the rejection criterion, it is only the location of considered alternatives which determines the location of the rejection region. If no alternative is specified, then there is no basis for preferring any one equal-sized region over another as the area of rejection. For a test of size .05, one might as well decide on the basis of rolling a regular icosahedron and dispense with the expensive apparatus of an experiment.

The existence of alternative hypotheses then implies the possibility of a second kind of error, which technically became known as an error of the second kind, or Type II. The concept had been used in the 1928 paper, but was not formally dealt with and named until 1933. In fact, in that paper (Neyman & Pearson, 1933b) they formally allowed a second partition of the sample space, creating a “region of doubt,” leading neither to acceptance nor rejection, but it was accorded a secondary importance and was never pursued.

There is in general no way of jointly minimizing the two types of errors. The solution of Neyman and Pearson is well known: to fix the probability of a Type I error in advance, then, insofar as possible, to choose a region so as to minimize the probability of a Type II error for the classes of alternatives entertained. The centerpiece of the joint theory is the Neyman-Pearson Lemma, according to which, in testing a simple hypothesis against a simple (point) alternative, the best test, in the sense of maximum power for fixed size, is a likelihood ratio test (Neyman & Pearson, 1933a). For composite alternatives specifying only a range of values for the parameter, “uniformly most powerful” tests could still be defined by this method when the parent distribution was normal. Less well-behaved distributions, on the other hand, may generate a profusion of choices, which the investigator is ultimately left on his or her own to resolve; and Neyman, in the last sentence of his paper (1937a), “emphasizes the rareness of cases where there exists a uniformly most powerful test” (p. 380).

A final consequence of the frequentist orientation is that the reasoning involved in statistical inference—or decision—is wholly hypothetical. It can be exhibited as a “relaxed” form of modus tollens: If H_0 is true, results X are unlikely; (some) X is observed; therefore, we decide, H_0 is false. Partly as a consequence of the historical

emphasis, under Fisher's influence, on significance levels at the expense of power, research workers have tended to lose sight of the hypothetical character of the process. As Pollard and Richardson (1987) have discussed very clearly, the conditional assertion, "If H_0 is true, there is only a 5% chance of results like these," has tended to condense to the apodictic, "There is only a 5% chance of being mistaken." Hence the jocular characterization of statisticians as those whose aim in life is to be wrong 5% of the time. What the significance level represents, however, is the proportion of false rejections of a true null in repeated tests on the same hypothesis, *not* the long-run proportion of Type I errors we shall make in all tests conducted at the .05 level. That value depends on what proportion of hypotheses we test are actually true. If all hypotheses we tested (at the .05 level) were true, then we should indeed be wrong (make a Type I error) in 5% of our decisions, though we should be wrong in 100% of our rejections. At the opposite extreme, if none of the hypotheses we tested were true, we would make no Type I errors (this is essentially the case in psychology, as Oakes, 1986, points out), and 100% of our rejections would be correct. The same considerations apply to power as to significance levels. In no case is it possible to know what proportion of our decisions—of our rejections or acceptances—should be correct unless we know what proportion of the hypotheses we test are true. And if we were in a position to know what proportion of the hypotheses we tested were correct, we should probably not be testing them anyway.

The issue can be conceived, as Spielman (1973) structures it, as a matter of base rates of true hypotheses in the reference set of hypotheses tested. Seen in this way, it is similar to the problem of antecedent probabilities and cutting scores discussed by Meehl and Rosen (1955). If fewer than 1 in 20 of the hypotheses we tested were false—in other words, if the prior probability of an hypothesis being true were more than .95—we should be better off accepting them all and dispensing with the tests. If, as is more nearly the case in psychological research, practically all hypotheses tested were false, then testing, relative to across-the-board rejection, results in more Type II errors. In the absence of knowledge of the incidence of true hypotheses among those tested, size and power are not necessarily useful features of a test; we cannot know, in fact, how useful they may be. Spielman's conclusion is similar to Hacking's:

Low size and high power are desirable if one is concerned solely with the average or overall performance of a test. However, once an experiment is performed, and a decision that really counts has to be made, the average performance of a test is irrelevant to determining what course of action is best to take. (1973, p. 211)

Spielman claims thus to have "refuted" the Neyman-Pearson theory, but Graves (1978) is right that he has done nothing of the kind. In fact, Neyman had considered the same problem many years ago. Having raised the question, "Without the a priori probabilities being known, is the theory able to say how frequently shall we be wrong or correct in applying this or that test?" he replies:

Of course the answer . . . is negative. To give an affirmative answer to this question would be equivalent to a denial of the correctness of Bayes' formula. Without knowing how frequently we shall have to deal with any of the admissible hypotheses, we are not able to say

how frequently we shall be right in dealing with them according to this or that rule. This may be considered unfortunate, but we cannot help it, and have simply to face the fact. (Neyman, 1942, p. 319)

Thus Neyman essentially accepted Spielman's "refutation" 30 years before it was written, and Spielman cannot really claim to have refuted the theory on its own terms, as he intended to do; his reply is only that if Neyman-Pearson theory is concerned exclusively with low long-run cost of erroneous decisions, and not with reliability of tests in his sense, then the theory is so restricted in scope as not to be worth the bother of refuting.

7.3.2 Confidence Intervals

Historically, confidence intervals have tended to appear as Neyman and Pearson's answer to fiducial probability. Though the appearance is not grievously wrong, in fact the original ideas, according to Neyman, were his own, developed while he was in Poland, and before he had any knowledge of Fisher's work on fiducial probability. "The term 'confidence interval' is a translation of the original Polish 'przedział ufności'" (Neyman, 1941, p. 128). The first presentation of the ideas in English was in an appendix to Neyman's (1934) paper "On the Two Different Aspects of the Representative Method" and an article by Clopper and Pearson (1934), which appeared at the same time in *Biometrika*, entitled "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial." As the text and the title of the latter article itself make clear, the authors initially took confidence coefficients and fiducial probabilities to be the same thing, and introduced the new term (Neyman's) only because they saw trouble in Fisher's notion of a fiducial probability distribution. Neyman and Pearson manipulated the same inequality that Fisher did to obtain their confidence interval, but their approach was distinguished right from the start by their insistence that the random variable was the interval itself rather than the parameter being estimated.

We cannot therefore say that for any specified value of x [i.e., the sample statistic; in this case, the number of successes in a binomial experiment] the probability that the confidence interval will include p is .95 or more. The probability must be associated with the whole belt, that is to say with the result of the continued application of a method of procedure to all values of x met with in our statistical experience. (Clopper & Pearson, 1934, pp. 407–408)

They go on to say:

Indeed it will be clear that if we had information a priori regarding the values of p likely to be met in our experience, and if this information could be expressed in precise numerical form, it would be possible to shift the confidence belt and so narrow the limits of uncertainty while retaining the same risk of error. For instance, if we knew that $1/3 \leq p \leq 2/3$, we should certainly cut off the two points of the lozenge by lines at $p = 1/3$ and $p = 2/3$.

In practice, however, it is rare

- (1) for the a priori information to be expressed in exact form,
- (2) even when it appears so expressible, for the working statistician to have time to calculate suitable modification for the limits.

Under general conditions, therefore, the statistician will usually be satisfied with limits which are “safe” in the sense that they give an expectation of long run accuracy which is precisely known, and thus avoid the uncertain risk of error involved in an attempt to introduce a priori information. (pp. 408–409)

Four related features of confidence intervals call for comment, and each has its counterpart in significance testing: the problem of prior knowledge, the orientation to long-run average performance, advance specification of risk, and the meaning of a confidence statement.

In the first place, Clopper and Pearson justify the neglect of prior information as the “safe” policy, in that it avoids an “uncertain risk of error.” But Kyburg’s (1963b) protest about fiducial probability seems relevant here as well: “To make a mind empty of background knowledge a condition of statistical inference is like making total damnation a necessary condition of salvation” (p. 939). Kendall (1949) proposes a hypothetical example:

Suppose I assume that a sampling process is such as to reproduce a binomial distribution—there is a good deal of evidence for this in the case of births. I observe a value of 0.60 as the ratio of male to total births in a sample of 10,000. The theory of confidence intervals says that I may assert that the proportion p lies between 0.59 and 0.61 with the probability that, if I make this type of assertion systematically in all similar cases, I shall be about 95 per cent right in the long run. But I do not then make such an assertion because I know too much about birth-rates to believe any such thing. The theory of confidence intervals gives no place to prior knowledge of the situation. How, then, can that theory provide a guide to conduct in making decisions? (p. 113)

If Kendall’s example seems unrealistic, consider this one from Oakes (1986):

Suppose . . . I wish to estimate the mean height of students at Birmingham University and take a random sample of size 2, recording observations of 73in and 73.5in. The 95% confidence interval for the mean height of Birmingham students can be calculated to be $(70.05\text{in} < \mu < 76.45\text{in})$. Faced with this interval presumably very few investigators would indeed be “95% confident” that the population value had been included in the interval. (p. 128)

The problem of neglected prior information may be seen as resulting from a characteristic of confidence intervals which they share with Neyman-Pearson tests of hypotheses: their orientation exclusively to long-run average performance. Individuating information about particular cases vitiates the statistical inference, which is based only on random aggregates. Critics of confidence interval theory have exploited this property to produce examples, not all of them “artificial,” where the theory leads to absurd results.

The following problem has occurred in several industrial quality control situations. A device will operate without failure for a time θ because of a protective chemical inhibitor injected into it; but at time θ the supply of this chemical is exhausted, and failures then commence, following the exponential failure law. It is not feasible to observe the depletion of this inhibitor directly; one can observe only the resulting failures. From data on actual failure times, estimate the time θ of guaranteed safe operation by a confidence interval. (E. T. Jaynes, 1976, p. 196)

Suppose 3 observations are made, with values of 12, 14, and 16. Then the shortest 90% confidence interval turns out to be $12.1471 < \theta < 13.8264$. But this interval for θ , the time of guaranteed safe operation, lies entirely beyond the time of the first failure!

Part of the problem is that, although it was never part of official doctrine that confidence intervals be based on sufficient statistics, in practice, when they are not (as in Jaynes' example),

It is possible to find a “bad” subclass of samples, *recognizable from the sample*, in which use of the confidence interval would lead us to an incorrect statement more frequently than is indicated by the confidence level; and also a recognizable “good” subclass in which the confidence interval is wider than it needs to be for the stated confidence level. The point is not that confidence intervals fail to do what is claimed for them; the point is that, if the confidence interval is not based on a sufficient statistic, it is possible to do better in the individual case by taking into account evidence from the sample that the confidence interval method throws away. (Jaynes, 1976, p. 199)

Whereas the preceding criticisms challenge the theory of confidence intervals ab extra, Hogben (1957) takes it to task on its own terms, in claiming it fails to meet consistently the criteria of the frequency theory. What Hogben felicitously calls the Forward Look requires that all criteria for inference be stated in advance; if the kind of statement we shall make—for instance, the precision of an estimate—can only be determined after the fact, we have lapsed into the Backward Look of Bayes and Laplace—and Fisher. But the binomial example of Clopper and Pearson fails on just this account. A strictly Forward Look, Hogben contends, would entail that a sample size could be calculated to guarantee a fixed risk (e.g., $\alpha = .05$) associated with an interval of a specified width (e.g., $P \pm .1$, where P is the sample proportion). In the binomial distribution, however, the variance depends on the unknown p ; if P is used as an estimate, then the width of the interval varies from sample to sample. Hence Clopper and Pearson's “lozenge” in place of a confidence belt of fixed width. A logical requirement, for Hogben, is then that all confidence interval statements must be made with respect to the greatest width of the belt. A consequence of this restriction, however, is that stochastic induction is sometimes less informative than non-stochastic induction based on the same data. As the extreme example, if all observations were of one kind (all successes), we should be worse off adopting the stochastic approach of confidence intervals, with a uniformly wide belt, than in drawing the obvious nonstochastic inference that $p = 1$. Jaynes' example is another instance where confidence intervals are less informative than nonstochastic inference.

Hogben's objection is scarcely confined to the binomial or exponential distributions. It obviously invalidates, to the same degree, confidence intervals based on the t distribution, where the width is similarly dependent on the sample standard deviation, as well as intervals for σ^2 based on the χ^2 distribution, or for a set of means based on the F distribution. Hogben, in fact, sees small-sample theory in its entirety as compromised by the Backward Look. In the more than half century since he put forth his objections from a strict behaviorist viewpoint, no one, to my knowledge, has either replied to them or heeded them.

There is, however, one major feature of confidence interval theory which has troubled even the writers of some statistics textbooks: the meaning of a confidence interval statement. Neyman describes it as follows. Suppose that some parameter θ_1 is being estimated by means of confidence intervals, and let E represent the sample point or event, with E' indicating the particular values of the x 's obtained in the present case. Of the confidence limits $\underline{\theta}(E)$ and $\bar{\theta}(E)$, Neyman (1941) says:

Their use in practice would consist of (i) observing the value E' of the x 's, (ii) calculating the corresponding values of the confidence limits $\underline{\theta}(E')$ and $\bar{\theta}(E')$, and (iii) *stating* that the true value θ_1 of θ_1 lies between $\underline{\theta}(E')$ and $\bar{\theta}(E')$. The justification is simple and perfectly in line with the classical point of view of probability: in many applications, the relative frequency of cases in which the statement $\underline{\theta}(E) \leq \theta_1 \leq \bar{\theta}(E)$ is correct will be approximately equal to $\alpha = 0.99$, whether or not the parameters for estimation are the same in all cases.

The word "stating" above is put in italics to emphasize that it is not suggested that we can "conclude" that $\underline{\theta}(E') \leq \theta_1 \leq \bar{\theta}(E')$, nor that we should "believe" that θ_1 is actually between $\underline{\theta}(E)$ and $\bar{\theta}(E)$. In the author's opinion, the word "conclude" has been wrongly used in that part of statistical literature dealing with what has been termed "inductive reasoning." Moreover, the expression "inductive reasoning" itself seems to involve a contradictory adjective. The word "reasoning" generally seems to denote the mental process leading to knowledge. As such, it can only be deductive. Therefore, the description "inductive" seems to exclude both the "reasoning" and also its final step, the "conclusion." If we wish to use the word "inductive" to describe the results of statistical inquiries, then we should apply it to "behaviour" and not to "reasoning." (p. 132)

Elsewhere he says:

The theoretical statistician constructing the functions $\underline{\theta}(E)$ and $\bar{\theta}(E)$. . . may be compared with the organizer of a game of chance in which the gambler has a certain range of possibilities to choose from while, whatever he actually chooses, the probability of his winning and thus the probability of the bank losing has permanently the same value, $1 - \alpha$.

The choice of the gambler on what to bet, which is beyond the control of the bank, corresponds to the uncontrolled possibilities of θ_1 having this or that value. The case in which the bank wins the game corresponds to the correct statement of the actual value of θ_1 . In both cases the frequency of "successes" in a long series of future "games" is approximately known. On the other hand, if the owner of the bank, say, in the case of roulette, knows that in a particular game the ball has stopped at the sector No. 1, this information does not help him in any way to guess how the gamblers have betted. Similarly, once the sample E' is drawn and values of $\underline{\theta}(E')$ and $\bar{\theta}(E')$ determined, the calculus of probability adopted here is helpless to provide answer [sic] to the question of what is the true value of θ_1 . (Neyman, 1937a, pp. 349–350)

The delicacy of confidence interval theory is that, while the algebraic statement $p(M - t_{.025}s/\sqrt{N} \leq \mu \leq M + t_{.025}s/\sqrt{N}) = .95$ is true, the arithmetic statement, for example $p(0.6 \leq \mu \leq 0.8) = .95$, is meaningless: The population mean either lies between 0.6 and 0.8, or it does not, and we don't know which; but there is no probability distribution about it. The random variables in the algebraic inequality are the limits of the interval, rather than μ ; the arithmetic inequality has no random variable to serve as the object of a probability statement. But the distinction between treating the interval limits and the parameter as having a probability distribution is an extremely subtle one and is lost on nearly all practitioners of the theory. Even an article (Natrella, 1960) purporting to explain confidence intervals to behavioral

scientists missed the point and spoke of the probability that a population mean lay between 0.6 and 0.8. Says Gridgeman (1963):

The average research worker, not gripped in the strait jacket of the frequency theory of probability, never has qualms about a hypothetical distribution of conceivable values of a parameter. But that same worker, in my experience, is often repelled by, and indeed apt to become ribald about, Neyman's stern advice to "carry out your experiments, calculate the confidence interval, and *state* that [the unknown parameter] belongs to this interval. If you are asked whether you 'believe' that [it] belongs to the confidence interval, you must refuse to answer." (p. 396)

The research worker is not entirely to blame. The frequency theory of probability prescinds utterly from haecceity, so that the 95% probability statement pertains theoretically to *any* interval but somehow never *this* one.

On an empirical-objectivist or relative frequency interpretation of probability, we will never be able to use such locutions as "the probability is 0.95 that the parameter lies in the interval I." We are prohibited from using such locutions no less than, and just as obviously as, we are prohibited from making probability statements about the weather *tomorrow*, the next toss of the coin, the frequency of heads on the *next 100* tosses, the life expectancy of *Mr. Jones*, etc. It is, as all frequentists have pointed out, quite beside the point to argue that people do in fact use these locutions. People often use "velocity" for speed, rather than for a speed-direction vector, but that does not prevent velocity from being a far more fruitful concept in mechanics than speed. But it is to the point that frequentists talk this way too, in ordinary life—most of them; and that the occasions when probability talk is appropriate become narrowed down to the point where "properly speaking" the only context in which it is appropriate to talk of probabilities is that of a statistical hypothesis. Something must fill the gap, and does. It is sometimes "significance level"; sometimes "confidence level." The most careful of statisticians (e.g., Neyman) may succeed in never using these concepts to characterize *particular* experiments (for it is the experimental test *in general*, not the particular test applied to a particular sample, to which these phrases may be appropriately applied), but we all know perfectly well that practically every statistician, except when he is teaching classes, uses these concepts in much the way that ordinary people use "probability." (Kyburg Jr., 1974, p. 57)

Pratt (1961) puts the matter succinctly: On the question of whether the confidence level, once we have inserted numerical values for the endpoints of the interval, measures the probability that the interval contains the parameter, he says: "We think, and would like to say, it 'probably' does; we can invent something else to say, but nothing else to think" (p. 165).

7.4 Differences Between the Fisher and Neyman-Pearson Theories

The Neyman-Pearson theory differs from Fisher's in several respects, some of which have already been indicated.

There is first of all a difference in terminology: The term "null hypothesis" was introduced by Fisher (1935a, pp. 18–20), and was not used by Neyman and Pearson. Fisher used the term to designate the hypothesis tested. He is usually interpreted as

having meant it in the sense of the hypothesis to be nullified; others (e.g., Lindquist, 1940) have taken it as the specification of a zero value for a parameter. Fisher argued, in any case, that the important point was that the null hypothesis specify *some exact* value of a parameter. If that condition was met, then the investigator could designate any hypothesis as the null; but in most research, the alternative to a “null” value was only a diffuse range, which could not be used as the basis of a probability distribution.

Neyman and Pearson, in contrast to Fisher, argued for the symmetry of the hypothesis tested and its denial, and contended that designation of some value as the hypothesis tested is in principle arbitrary, to be determined practically by the nature of the problem.

In fact, an error which is “of the first kind” when we test some hypothesis H_0 becomes “of the second kind” when we test its negation, say \bar{H}_0 , and conversely. The distinction in question is connected with the fact that in various problems of application the importance of one kind of error, whether we call it first or second, is by far greater than that of the other, and this must be taken into account when choosing the critical region. . . .

A more or less general convention was adopted to consider as the hypothesis tested the one by which the errors of the first kind are of greater importance than those of the second. Thus in the... example of testing drugs, the hypothesis tested would be H_0 : “the actual toxicity of the drug does exceed the prescribed safety limit,” and the error of the first kind would be committed if (i) this hypothesis is in fact correct and (ii) if the test rejects it. (Neyman, 1942, p. 304)

In the second place, as has been noted before, Fisher never granted the necessity of specifying alternatives in tests of hypotheses and spoke rather scornfully on occasion of the doctrine of Type I and Type II errors. His usual defense was simply a reiteration of the significance testing procedure as he saw it, but he also used as an argument for his position the chi-square test of goodness of fit. This was among the oldest tests, and the established practice had not included the specification of any particular form of departure from, say, normality. But Neyman and Pearson were scarcely oblivious to the problem; each of them published articles (Neyman, 1937b/1967a; Pearson, 1942) showing how selection of an optimal test criterion in such tests depends on the particular form of departure from the hypothesized distribution to which we want the test to be sensitive, and studying the power of goodness-of-fit tests under various kinds of alternatives. In any case, they surely had a sufficient rebuttal in the claim that, without reference to alternatives, no justification could be given for choosing the tails of the sampling distribution as the rejection region over any other set of size α . Extreme improbability of an event on the basis of chance—a very low α —does not in itself warrant rejection of “chance” as an explanation, as Gosset recognized in his example of dealing himself 13 trumps.

Related to the foregoing point is Fisher’s contention that in a statistical test we either reject the null hypothesis or fail to reject it:

It should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (Fisher, 1935a, p. 19)

(His speaking here of proof and disproof is curious, inasmuch as statistical methods do not allow the possibility of proof or disproof as the terms are ordinarily understood.) This principle has been widely adopted by psychologists; it is often said that acceptance of the null hypothesis implies the demonstration of an absence, whereas epistemologically the onus of proof is on them who assert the positive (i.e., the alternative hypothesis). Neyman (1942), however, explicitly equated acceptance with failure to reject, and not to do so is in fact incompatible with his theory; for, as Rozeboom (1960) and others have pointed out, unless we are prepared to accept the null hypothesis if it is not rejected, there can be no possibility of a Type II error—in which case we ought never to reject the null and thus avoid all errors of both kinds.

As noted above, the frequency theory logically commits us to advance specification of the size of the rejection region; otherwise the probabilities of Type I errors are uncontrolled, and the significance level, understood as a proportion of false rejections of the null in a long series of formally similar tests, carries no meaning. Fisher gave signs of appreciating the point; in *The Design of Experiments*, he wrote, “In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure” (1935a, p. 16). The view expressed here is of course in perfect accord with that of Neyman and Pearson. Elsewhere, however, even in the same book, Fisher was oriented more to the individual test than to long sequences of them, and, as an important result, he allowed the specification of the rejection region after the data have been collected and analyzed.

Interestingly, Fisher’s ambivalence has been preserved in psychological research. Some of the more sophisticated textbooks of statistics for psychologists (e.g., Hays, 1963, 1981) are careful to state that the significance level must be chosen in advance, even if they do not say very much about the reason; but many, especially on the elementary level, leave the impression, if they do not actually say, that the investigator may always designate post facto the smallest possible region which would have included the observed result, as its level of significance; and the practice of appending from one to three asterisks after the reported value of a test statistic indicates that the less defensible side of Fisher’s ambivalence has been accepted widely enough to have become a convention. Bhattacharyya and Johnson (1977), for example, call the Fisherian, post facto probability the *significance probability*, to distinguish it from the *significance level*, α , which is specified in advance.

Underlying these differences is some disagreement on the meaning of a significance level. In spite of his professed adherence to a frequency theory of probability, Fisher wanted to attach some meaning to probability statements about the results of individual experiments, and he regularly denigrated the approach which identified hypothesis testing with acceptance sampling.

The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who “rejects” a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and

when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. It should not be forgotten that the cases chosen for applying a test are manifestly a highly selected set, and that the conditions of selection cannot be specified even for a single worker; nor that in the argument used it would clearly be illegitimate for one to choose the actual level of significance indicated by a particular trial as though it were his lifelong habit to use just this level. Further, the calculation is based solely on a hypothesis, which, in the light of evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be, for this reason only, much less than the frequency specifying the level of significance. A test of significance contains no criterion for “accepting” a hypothesis. According to circumstances it may or may not influence its acceptability. (Fisher, 1956/1973, pp. 44–45)

Fisher's refusal to identify the significance level with the long-run frequency of false rejections of the null hypothesis may come as something of a surprise, since textbooks for research workers generally present them as indistinguishable. To drive a wedge between them, and thus to demonstrate, in his view, the superiority of his approach, Fisher relied principally upon the famous Behrens-Fisher problem. This is simply the problem of testing the difference between means of two normal populations with unequal variances; a solution was first given by W.-V. Behrens in 1929. Fisher (1935b) derived an equivalent solution by the fiducial argument, to which no counterpart could be produced in the Neyman-Pearson theory. The Behrens-Fisher solution turned out to rest on the assumption of a particular fiducial distribution for the ratio of the two standard deviations; without some such assumption, the problem could not be solved. Thus Fisher had a problem for which he could compute a significance level, but Neyman-Pearson theory could not produce a long-run relative frequency of false rejections. The practical problem was solved by Welch (1938), via nonintegral adjustment of the degrees of freedom for the Student t ; as a result, the Behrens-Fisher problem has remained mostly unknown to psychologists (though a few textbooks, like Howell, 1982, 1987, refer to it by that name).

The differing interpretations of significance levels reflect, again, the underlying difference in whether the population is regarded as real or hypothetical. Paradoxically, it is Fisher, concerned with the real problems of research, for whom the population was a hypothetical invention, and Neyman, the abstract mathematician, for whom it was a concrete reality. But what was real in Fisher's research—and those agricultural and social scientists following him—was only his few observations. They cannot be called a sample, because there never was really any population. Fisher, as we have seen, was clear—cavalier, it could be said—about the hypothetical infinite population being an invention needed only to allow the mathematics of probability theory to be brought to bear: “The notion of repeated sampling from a fixed population has served its sole purpose when the distribution of t has been established” (Fisher, 1941, p. 148).

The circularity of the construct which Kendall (1943) and Hacking (1965) objected to as a problem, however, Fisher pushed as an advantage. In particular, it allowed him to argue in several notable cases that the population should be defined

to match particular characteristics of a given sample. At least two of these were taken as the standard solutions in psychological research.

One was the distribution of the regression coefficient. If the X s as well as the Y s are assumed to vary from sample to sample, the distribution problem is difficult. By assuming the X s fixed, however, Fisher (1922a) was able to show rather easily that b has the t distribution. He argued that the relevant population from which we could consider ourselves to have sampled was precisely one in which all the X s were the same as in our data, and that argument carried over into the textbooks (e.g., Hays, 1963, 1971). Bartlett (1938) soon proved that the regression coefficient was also distributed as t when the joint distribution of X and Y was multivariate normal; the textbooks mostly seem to have taken this result to mean that it didn't matter what the distribution of X was.

Another, similar case did lead to more controversy: the issue of marginal totals in contingency tables. The mathematics for the situations of one or both sets of marginals being variable is not difficult, and Egon Pearson (1947) published an exploration of the various cases. Although in most applications in psychological research the model of variable marginals would appear to be more appropriate, Fisher's solution, assuming that all possible repetitions of the experiment would yield the same marginals, is all that has survived.

Notwithstanding these important applications where a Fisherian approach prevailed, the fundamental difference between Fisher and Neyman and Pearson with respect to their goals—to analyze data or to formulate a good statistical theory—led, not so much to different numerical results, as to different interpretations and applications. Fisher stated the meaning, or relevance, of a significance level, not just in epistemological, but in explicitly psychological terms. At the same time, of course, he dissociated himself from the view of probability as subjective and issued a reminder that significance levels must be understood as direct, rather than inverse, probabilities:

Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders. It is more primitive, or elemental than, and does not justify, any exact probability statement about the proposition. . . .

The psychological resistance has been, I think wrongly, ascribed to the fact that the event in question *has*, in the proper sense of the Theory of Probability, the low probability assigned to it, rather than to the fact, very near in this case, that the correctness of the assertion would *entail* an event of this low probability. The probability statement is a sufficient, but not a necessary, condition for disbelief in this degree.¹¹ (Fisher, 1956/1973, pp. 46–47)

¹¹He went on to say:

Disbelief is equally justified when the probability is hypothetical. . . .

In general, tests of significance are based on *hypothetical* probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test. (p. 47)

Neyman and Pearson, as we have seen, initially took similar aims, in situating their theory in problems of inference, but the logic of their theory brought them inexorably to emphasize applications in manufacturing and other situations where repeated sampling was more of a reality than in field or laboratory research. Neyman's emphasis on "inductive behavior," in contradistinction to inductive inference, did not always add to the attractiveness of the theory for those seeking a theory of inference, although it made him an instant winner with American psychology, which was at its most behaviorist at that time. Fisher may never have succeeded in making his theory *consistent* with his epistemological intentions, but his description of what he was doing sounded more relevant than Neyman and Pearson's. Kendall (1963) summarizes the dispute as follows:

The position on both sides has been restated *ad nauseam*, without much attempt at reconciliation or, as I think, without an explicit recognition of the real point, which is that a man's attitude towards inference, like his attitude towards religion, is determined by his emotional make-up, not by reason or by mathematics.

More recently the ground of controversy has moved still farther, though the basic point remains the same. Under the influence of Neyman and Wald (who also came under the Fisherian anathema, though posthumously), there has been a strong movement in the U.S.A. to regard inference as a branch of decision theory. Fisher would have maintained (and in my opinion rightly) that inference in science is not a matter of decision, and that, in any case, criteria for choice in decision based on pay-offs of one kind or another are not available. This, broadly speaking, is the English as against the American point of view. We shall see a lot of water under the bridge before this conflict is resolved. Not wishing to be controversial about it, I propound the thesis that some such difference of attitude is inevitable between countries where what a man does is more important than what he thinks, and those where what he thinks is more important than what he does. (p. 4)

The handling of the pervasive tension between interpretations by authors of textbooks for psychologists is discussed in Chap. 9.

References

- Bacon, F. (1627). *Sylva sylvarum, or a naturall history in ten centuries*. London, UK: William Rawley.
- Baird, D. (1982). Significance tests: Their logic and early history (Doctoral dissertation, Stanford University, 1981). *Dissertation Abstracts International*, 42, 3629A. (University Microfilms No. 8202056)
- Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. *Mathematical Proceedings of the Cambridge Philosophical Society*, 34, 33–40.
- Bhattacharyya, G. K., & Johnson, R. A. (1977). *Statistical concepts and methods*. New York, NY: Wiley.
- Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York, NY: Wiley.
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413.
- Cushny, A. R., & Peebles, A. R. (1905). The actions of optical isomers. II. Hyoscines. *Journal of Physiology*, 32, 501–510.
- Doll, R., & Hill, A. B. (1950, September 30). Smoking and carcinoma of the lung: Preliminary report. *British Medical Journal*, 2, 739–748.
- Edwards, A. W. F. (1972). *Likelihood*. Cambridge, UK: Cambridge University Press.

- Edwards, A. W. F. (1977). Discussion of Wilkinson's paper. *Journal of the Royal Statistical Society*, 398, 144–145.
- Enlow, E. R. (1937). *Statistics in education and psychology*. New York, NY: Prentice-Hall.
- Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41, 155–160. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 1, pp. 53–58).
- Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507–521.
- Fisher, R. A. (1922a). The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85, 597–612.
- Fisher, R. A. (1922b). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (Series A)*, 222, 309–368.
- Fisher, R. A. (1924). The distribution of the partial correlation coefficient. *Metron*, 3, 329–332. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 1, pp. 489–492).
- Fisher, R. A. (1925a). Applications of "Student's" distribution. *Metron*, 5, 90–104. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 2, pp. 41–55).
- Fisher, R. A. (1925b). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver and Boyd. (4th ed., 1932).
- Fisher, R. A. (1925c). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26, 528–535.
- Fisher, R. A. (1933). The contributions of Rothamsted to the development of the science of statistics. *Annual Report of the Rothamsted Experimental Station*, 43–50. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 3, pp. 84–91).
- Fisher, R. A. (1935a). *The design of experiments*. Edinburgh, Scotland: Oliver and Boyd.
- Fisher, R. A. (1935b). The fiducial argument in statistical inference. *Annals of Eugenics*, 6, 391–398. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 3, pp. 317–324).
- Fisher, R. A. (1936a). "The coefficient of racial likeness" and the future of craniometry. *Journal of the Royal Anthropological Institute*, 66, 57–63.
- Fisher, R. A. (1936b). Uncertain inference. *Proceedings of the American Academy of Arts and Science*, 71, 245–258.
- Fisher, R. A. (1941). The asymptotic approach to Behrens's integral, with further tables for the d test of significance. *Annals of Eugenics*, 11, 141–172. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 4, pp. 323–354).
- Fisher, R. A. (1945). The logical inversion of the notion of the random variable. *Sankhya*, 7, 129–132.
- Fisher, R. A. (1951). Statistics. In A. E. Heath (Ed.), *Scientific thought in the twentieth century* (pp. 31–55). London, UK: Watts. (Reprinted in J. H. Bennett, Ed., *Collected papers*, Vol. 5, pp. 185–207).
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). New York, NY: Hafner. (1st ed., 1956).
- Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311–320.
- Fisher, R. A., & Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. Edinburgh, Scotland: Oliver and Boyd.
- Fraser, D. A. S. (1963). On the definition of fiducial probability. *Bulletin of the International Statistical Institute*, 40, 842–856.
- Fraser, D. A. S. (1968). *The structure of inference*. New York: Wiley.
- Graves, S. (1978). On the Neyman-Pearson theory of testing. *British Journal for the Philosophy of Science*, 29, 1–23.
- Grigdeman, N. T. (1963). Discussion, Conference on Fiducial Probability. *Bulletin of the International Statistical Institute*, 40, 936–937.

- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Halbwachs, M. (1972). La statistique en sociologie [Statistics in sociology]. In V. Karady (Ed.), *Classes sociales et morphologie* [Social classes and morphology] (pp. 329–348). Paris, France: Minuit. (Original work published 1935).
- Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19, 491–544.
- Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart, Winston.
- Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, Winston.
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart, Winston.
- Hays, W. L. (1988). *Statistics* (4 ed.). New York: Holt, Rinehart, Winston.
- Hogben, L. (1957). *Statistical theory: The relationship of probability, credibility and error*. New York, NY: Norton.
- Hotelling, H. (1951). The impact of R. A. Fisher on statistics. *Journal of the American Statistical Association*, 46, 35–46.
- Howell, D. C. (1982). *Statistical methods for psychology*. Boston, MA: Duxbury Press. (2nd ed., 1987).
- Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science. Vol. 2. Foundations and philosophy of statistical inference* (pp. 175–258). Dordrecht, The Netherlands: Reidel.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press. (1st ed., 1939).
- Kealey, T. (2021). Terence Kealey on the myths of public funding of science. *The Accad and Koka Report*, Episode 159.
- Kempthorne, O. (1976). Statistics and the philosophers. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science. Vol. 2. Foundations and philosophy of statistical inference* (pp. 273–314). Dordrecht, The Netherlands: Reidel.
- Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika*, 56, 231–248.
- Kendall, M. G. (1942). On the future of statistics. *Journal of the Royal Statistical Society, Series B*, 105, 69–80.
- Kendall, M. G. (1943). *The advanced theory of statistics* (Vol. 1). New York, NY: Lippincott.
- Kendall, M. G. (1949). Reconciliation of theories of probability. *Biometrika*, 36, 101–116.
- Kendall, M. G. (1961). Studies in the history of probability and statistics. XI. Daniel Bernoulli on maximum likelihood. *Biometrika*, 48, 1–18.
- Kendall, M. G. (1963). Ronald Aylmer Fisher, 1890–1962. *Biometrika*, 50, 1–15.
- Keynes, J. M. (1973). *A treatise on probability*. New York, NY: St. Martin's Press. (Original work published 1921).
- Kolmogorov, A. N. (1950). *Foundations of the theory of probability*. New York, NY: Chelsea. (Original work published 1933).
- Kyburg, H. E., Jr. (1963a). Logical and fiducial probability. *Bulletin of the International Statistical Institute*, 40, 884–901.
- Kyburg, H. E., Jr. (1963b). Discussion, Conference on Fiducial Probability. *Bulletin of the International Statistical Institute*, 40, 938–939.
- Kyburg, H. E., Jr. (1974). *The logical foundations of statistical inference*. Dordrecht, The Netherlands: Reidel.
- Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York, NY: Wiley.
- Lindley, D. V. (1963). Discussion, Conference on Fiducial Probability. *Bulletin of the International Statistical Institute*, 40, 919–921.
- Lindquist, E. F. (1940). *Statistical analysis in educational research*. Boston, MA: Houghton Mifflin.
- Lucas, J. R. (1970). *The concept of probability*. Oxford, UK: Clarendon Press.
- Mackenzie, D. (1981). *Statistics in Britain, 1865–1930: The social construction of scientific knowledge*. Edinburgh, Scotland: Edinburgh University Press.

- Mahalanobis, P. C. (1938). Professor Ronald Aylmer Fisher. *Sankhya*, 4, 265–272.
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge, UK: Cambridge University Press.
- Mayo, O. (2014). Fisher in Adelaide. *Biometrics*, 70, 266–269.
- Medin, D. L., & Thau, D. M. (1992). Theories, constraints, and cognition. In H. L. Pick Jr., P. van den Broek, & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 165–187). Washington, DC: American Psychological Association.
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.
- Mises, R. v. (1957). *Probability, statistics and truth* (2nd English ed.). New York, NY: Macmillan. (Original work published 1928).
- Natrella, M. G. (1960, February). The relation between confidence intervals and tests of significance. *American Statistician*, 14(1), 20–22, 38.
- Neyman, J. (1930). Contribution to the theory of certain test criteria. *Bulletin of the International Statistical Institute*, 24(Part 2), 44–86.
- Neyman, J. (1934). On the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97, 558–625.
- Neyman, J. (1937a). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London (Series A)*, 236, 333–380.
- Neyman, J. (1937b). “Smooth” test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20, 149–199. (Reprinted in *Early statistical papers*, pp. 291–319).
- Neyman, J. (1938). L'estimation statistique traitée comme un problème classique de probabilité [Statistical estimation treated as a classical problem of probability]. *Actualités Scientifiques et Industrielles*, No. 739, 25–57.
- Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, 32, 128–150.
- Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 105, 292–327.
- Neyman, J. (1950). *First course in probability and statistics*. New York, NY: Holt.
- Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. Washington, DC: U.S. Department of Agriculture Graduate School.
- Neyman, J. (1967a). *A selection of early statistical papers*. Berkeley, CA: University of California Press.
- Neyman, J. (1967b). R. A. Fisher (1890–1962): An appreciation. *Science*, 156, 1456–1460.
- Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society (Supplement)*, 2, 107–180.
- Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, 20A, 175–240.
- Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika*, 20A, 263–294.
- Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London (Series A)*, 231, 289–337.
- Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society*, 24, 492–510.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.
- Oberschall, A. (1987). The two empirical roots of social theory and the probability revolution. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution. Vol. 2. Ideas in the sciences* (pp. 103–131). Cambridge, MA: MIT Press.
- Pearson, E. S. (1925). Bayes' Theorem, examined in the light of experimental sampling. *Biometrika*, 17, 388–442.

- Pearson, E. S. (1939). "Student" as statistician. *Biometrika*, 30, 210–250.
- Pearson, E. S. (1942). Notes on testing statistical hypotheses. *Biometrika*, 32, 311–316.
- Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. *Biometrika*, 34, 139–167.
- Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics*, 33, 394–403.
- Pearson, K. (1892). *The grammar of science*. London, UK: Walter Scott.
- Pitman, E. J. G. (1963). Discussion, Conference on Fiducial Probability. *Bulletin of the International Statistical Institute*, 40, 932.
- Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin*, 102, 159–163.
- Pratt, J. W. (1961). [Review of E. Lehmann, *Testing statistical hypotheses*]. *Journal of the American Statistical Association*, 56, 163–167.
- Reid, C. (1982). *Neyman—From life*. New York, NY: Springer.
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416–428.
- Savage, L. J. (1961). The foundations of statistics reconsidered. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 575–586). Berkeley, CA: University of California Press.
- Savage, L. J. (1963). Discussion, Conference on Fiducial Probability. *Bulletin of the International Statistical Institute*, 40, 925–927.
- Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R. A. Fisher*. Dordrecht, The Netherlands: Reidel.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Sheynin, O. B. (1971). J. H. Lambert's work on probability. *Archive for the History of Exact Sciences*, 7, 244–256.
- Spielman, S. (1973). A refutation of the Neyman-Pearson theory of testing. *British Journal for the Philosophy of Science*, 24, 201–222.
- "Student". (1908). The probable error of a mean. *Biometrika*, 6, 1–25.
- Tankard, J. W., Jr. (1984). *The statistical pioneers*. Cambridge, MA: Schenkman.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362.
- Welch, B. L. (1968). *The biometricians and the development of statistics*. Leeds, UK: Leeds University Press.
- Wilkinson, G. N. (1977). On resolving the controversy in statistical inference. *Journal of the Royal Statistical Society (Series B)*, 39, 119–171.
- Yates, F. (1951). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *Journal of the American Statistical Association*, 46, 19–34.
- Yates, F., & Mather, K. (1963). Ronald Aylmer Fisher, 1890–1962. *Biographical Memoirs of Fellows of the Royal Society of London*, 9, 91–120. (Reprinted in *Collected papers of R. A. Fisher*, Vol. 1, pp. 23–52).
- Yule, G. U., & Kendall, M. G. (1937). *An introduction to the theory of statistics* (11th ed.). London: Griffin.

Chapter 8

Bayesian Theories of Probability and Statistical Inference



As was discussed in Chap. 3, the concept of probability was originally an epistemic one, onto which an aleatory meaning was grafted in the late seventeenth century. Through the eighteenth century, and much of the nineteenth, the concept remained implicitly dualistic and occasioned a degree of philosophical scrutiny which is to retrospect rather surprisingly slight. The emergence of hostile schools or theories of probability was instigated by criticism of the Laplacean theory of statistical inference, which was based on application of the principle of indifference to the values of unknown probabilities. The positivist rejection of that metaphysical sounding assumption by Venn, in favor of a frequency definition of probability, appeared to be the way to avoid the paradoxes of that theory and to place the subject on a modern empirical footing.

The reductionist frequency theory, however, was beset by its own paradoxes, as was discussed in Chap. 6. Were one writing history prospectively, one might have a theory which reaffirmed the original meaning of probability as fundamentally epistemic arising in response to the criticisms of Mises' theory. In fact, such a theory did not wait for Mises. One version, the so-called logical theory of probability, was developed more or less independently in the early decades of the twentieth century by John Maynard Keynes and Harold Jeffreys, though both were influenced by W. E. Johnson's lectures at Cambridge. Keynes published his *Treatise on Probability* in 1921, 18 years ahead of Jeffreys' *Theory of Probability*, but Jeffreys claims priority for an article (Wrinch & Jeffreys, 1919) published 2 years ahead of Keynes' book, which Keynes had not seen.

The other type of degrees-of-belief approach is the concept of personal probability, an idea which originated, again independently and at about the same time, with Frank Ramsey and Bruno de Finetti. The personalist theory of probability may also be viewed, interestingly enough, as a behavioral approach, for it takes an individual's betting quotients as the measure of probability—a psychological conception in contrast to the epistemological interpretation of Keynes and Jeffreys.

All of these theories are sometimes referred to collectively as “subjective” theories, to contrast them with “objective” frequency interpretations; but the epithet “subjective,” apart from being prejudicial, actually turns out to be off the mark, given the nature of even the personalist theory. The theories in question could be called “epistemic” to distinguish them from those based on “aleatory” probability alone, but the characterization fits the personalist theories much less well than the logical version. The term “Bayesian” is not satisfactory, either—first, because Bayes’ own conception was based on expectation rather than degrees of belief, and, second, because Bayes’ Theorem is independent of any interpretation of probability, and in fact finds application in Neyman-Pearson theory. “Degrees-of-belief theories” is perhaps the most accurate label, but it is also awkward and nonstandard. I shall use the term “Bayesian” here in view of its current acceptance, especially in statistical theory.

8.1 Logical Theories of Probability

8.1.1 Keynes

However much one may disparage Keynes’ contributions to economic theory and subsequent world crises, his *Treatise on Probability* must be acknowledged as an impressive book. He began work on it as a dissertation in 1906, when he was 23; though largely finished within 5 years, it was interrupted by other projects and was not published until 1921.¹ At the time Keynes was writing, the germs of a frequency theory were in the air, but the “bug” was still waiting to be caught, perhaps until the positivist philosophy of science had sufficiently weakened classical resistance to a reductionist program. Thus, though Keynes devoted a chapter in his book to discussion of “the frequency theory of probability,” he really had only Venn to discuss, and so his book represents only the second attempt in history at an integrated and comprehensive theory of probability. Since Venn’s theory he moreover dismissed, he was really starting afresh.

Keynes placed probability squarely in its traditional framework, as a part of epistemology concerned with the justification and formalization of argument, like the logic of implication, which in his vision would become a special case. He took probability to be “the degree of rational belief” in a proposition warranted by a particular body of evidence.

His definition of probability explicitly as a relation, between statement and evidence, might have seemed more natural two centuries earlier. Indeed he claims that

¹ Rothbard (1992), following Keynes’ biographer Skidelsky, thinks the *Treatise* was motivated by the desire to construct a justification for Keynes’ egocentrically antinomian twist on Moore’s ethics. I don’t know any more about Keynes’ motives, but that seems to me a stretch, in view of the small degree to which it would appear to have accomplished that aim, especially in proportion to the effort.

the first such definition was due to L. M. Kahle, in his *Elementa logicae Probabilium methodo mathematica in usum Scientiarum et Vitæ adornata*, published at Halle in 1735.² On Keynes' view, it is meaningless to speak of the probability of a proposition in the abstract, without reference to some body of data. If we do so in common usage, it is through ellipsis, when it is obvious which supporting data are alluded to. The probability of a proposition is thus subject to change with the acquisition of new relevant knowledge. Keynes introduced the virgule notation for conditionality (later modified by Jeffreys to the more familiar $|$, to avoid confusion with division): He wrote a/h to denote the probability of a on data h .

Keynes held that probability arises from gaps in our knowledge; hence there would presumably be no need for the concept were it not the case that our knowledge grows by degrees. But he was writing in advance of the great discoveries of twentieth-century physics, and critics have occasionally supposed his theory to have been invalidated automatically by the principle of indeterminacy and the notion that probability and chance may be objective on the level of elementary particles. Keynes actually considered the idea of objective probability, which had already been explored by Cournot (1843), but found it wanting. There may be phenomena, he said, which may meaningfully be said to be due to objective chance; but even there, he thought, scrutiny would reveal the logical interpretation as the more fundamental.

An event is due to objective chance if in order to predict it, or to prefer it to alternatives, at present equiprobable, with any high degree of probability, it would be necessary to know a great many more facts of existence about it than we actually do know, and if the addition of a wide knowledge of general principles would be of little use. (1921/1973, p. 289)

It is not altogether clear that Keynes' theory would be damaged beyond repair by the admission of some probabilities as, so to speak, metaphysical rather than epistemological: His concept of probability might still be defended, in other words, on the ground that gaps in reality leave gaps in our knowledge. Bateman (1987) notes that Keynes did accept the notion of objective chance in physics after 1938, but that he didn't see it as universally applicable.

The distinctive feature of a logical, as opposed to a personalist, theory is that probability represents not merely a degree of belief, but a degree of *rational* belief. Keynes was as much aware as anyone that people hold all sorts of convictions to a degree not at all commensurate with their actual warrant; but these variations, and only these, are the province of psychology. The probability of a proposition does vary from subject to subject, as each brings a unique evidential context to its assessment; but, given a particular body of evidence, there will be a particular degree to

² Keynes (1921/1973) remarked that "This work, which seems to have soon fallen into complete neglect and is now extremely rare, is full of interest and original thought" (p. 90n). His reference has sent more than one scholar off on a fruitless search. Kneale (1949) reports that he was unable to locate a copy, but the rarity of the book is more impressively attested to by the fact that neither was Isaac Todhunter (1865), whose history up through Laplace is considered encyclopedic.

which it *objectively* warrants the conclusion in question.³ That is what is meant in denying that probability is subjective, and it is the determination of that objective warrant which Keynes took as the primary task of probability theory.

Keynes' disavowal of subjectivism and psychologism has not been thought successful beyond dispute, however. Carnap (1962) mounted a particularly spirited attack, and one which is made the more interesting by the fact that Carnap's own theory of probability is often referred to as a "logical" theory. Carnap denied outright that logic or probability theory has anything to do with thinking or judgment; its subject matter instead is an inscrutable entity called "logical relations" (or probability relations). From that point of view, Keynes' theory represents ("qualified") psychologism. In a passage that is widely quoted (with obvious reason), Carnap writes:

The characterisation of logic in terms of correct or rational or justified belief is just as right but not more enlightening than to say that mineralogy tells us how to think correctly about minerals. The reference to thinking may just as well be dropped in both cases. Then we say simply: mineralogy makes statements about minerals, and logic makes statements about logical relations. The activity in any field of knowledge involves, of course, thinking. But this does not mean that thinking belongs to the subject matter of all fields. It belongs to the subject matter of psychology but not to that of logic any more than to that of mineralogy.⁴ (1962, pp. 41–42)

Carnap's criticism is weakened somewhat, however, by the fact that he himself was never able to articulate the ground of probability relations as he saw them, even to his own satisfaction.

Keynes' probability relations have also been criticized on the rather more obvious question of how they are known and measured. Kempthorne (1976), frustrated with the intractable abstractness of Keynes' theory, says that it is "like having a book on how to knit which merely talks of the theory of knitting" (p. 280). Keynes did not appreciably help his case, in the eyes of twentieth-century appraisers, by holding essentially that probability relations, being irreducible to other concepts, are ultimately "just known," by a kind of Aristotelian direct intuition. Perhaps the best that could be said in his defense is that Keynes only appears to hold a weak position because he offered inadequate answers to questions which more careful thinkers have simply dodged.

Consider the problem of measurement, where Keynes' answer cannot really be said to lack interest. In contrast to almost every other writer on the subject, Keynes

³ Keynes might also have gone on to say that to speak baldly of "the probability" of a proposition is to make tacit reference to the widest body of relevant evidence available to humankind at a given point in time. Just that idea was urged, in fact, by Borel (1924), in his review of Keynes' book.

⁴ In a precious commentary Toulmin (1964) adds:

[Carnap] treats all logical relations, and hence all justified beliefs, all evidential support and all satisfactory explanations as relying for their validity on considerations of semantics alone. Waismann has criticised Frege for thinking that the statements of logic represent "little hard crystals of logical truth": it is curious, therefore, that Carnap, following Frege, should put logical relations on a footing with minerals. (p. 87)

not only questioned whether all probabilities were numerically measurable but also denied that all probabilities are commensurable or orderable in magnitude. Comparison, he said, requires some dimension of communality, which may obtain in three types of cases:

- (a) a/hh_1 differs from a/h , where h_1 is “a single piece of information,” containing no independent parts which are relevant: The addition of new pertinent evidence may enable us to make a specifiable change in our probability estimate.
- (b) $ab/h \leq a/h$: The broader assertion of b together with a cannot be more probable than the assertion of a alone; and
- (c) combinations of (a) and (b), or cases based on the principle of indifference.

Otherwise, Keynes argued, probabilities may not be comparable. He suggested an analogy with similarities:

For instance, a book bound in blue morocco is more like a book bound in red morocco than if it were bound in blue calf; and a book bound in red calf is more like the book in red morocco than if it were in blue calf. But there may be no comparison between the degree of similarity which exists between books bound in red morocco and blue morocco, and that which exists between books bound in red morocco and red calf. (1921/1973, p. 36)

Similarly, there may be no basis for saying whether it is more probable, within the next century (on the vague, implicit body of available evidence) that we will enter another Ice Age, or (on a similarly vague body of evidence) that peace will be achieved in the Middle East. Any probability can be placed in an ordered series; but it may belong to more than one series, and a given series is not necessarily compact; i.e., it is not necessarily true that any two probabilities in a series have another probability between them. Keynes denied, in sum, that with respect to an arbitrary pair of probabilities, one of them is necessarily less than, equal to, or greater than the other.⁵ (A weak form of the assertion would be that we cannot always say what the relation is.) In a passage anticipative of Toulmin, he wrote:

⁵ It is of considerable interest that Keynes carried his skepticism about quantification over into economics, just because his *General Theory of Employment, Interest, and Money* was a tremendous stimulus to the growth of macroeconomics and econometrics. He made little use of mathematics in the *General Theory* (see Chap. 9), and a famous passage refers to:

symbolic pseudo-mathematical methods of formalizing a system of economic analysis... which allow the author to lose sight of the complexities and interdependencies of the real world in a maze of pretentious and unhelpful symbols. (quoted in Patinkin, 1976, p. 1093)

During the 1930s, Britain evidently lagged behind the United States in the collection of national income statistics, and Patinkin (1976) reproaches Keynes for not having exerted some of his very considerable influence toward that effort. But it seems clear that Keynes didn't place enough confidence in such data to value that endeavor very highly.

As skeptical as Keynes was about quantification and statistical analysis in economics, he was surely also aware that his skepticism about measurement of probabilities kept his *Treatise on Probability* from having any influence at all. So it is not surprising to find the *General Theory* more inconsistent on that issue, and his followers naturally ignored his caveats and seized on the analytic methods he ambivalently appeared to endorse.

Is our expectation of rain, when we start out for a walk, always *more* likely than not, or *less* likely than not, or *as* likely as not? I am prepared to argue that on some occasions, *none* of these alternatives hold, and that it will be an arbitrary matter to decide for or against the umbrella. If the barometer is high, but the clouds are black, it is not always rational that one should prevail over the other in our minds, or even that we should balance them,—though it will be rational to allow caprice to determine us and to waste no time on the debate. (p. 30)

Keynes' denial of strict orderability has been thought by most reviewers to cost him more than it was worth, particularly in terms of axiomatizability. But again, the only real alternative that has been offered, within a logical theory, is to exclude the troublesome cases from probability theory by definition. It might be proposed that all those probability comparisons of *practical interest* fall into one of the three categories above—after all, who but a modern philosopher or experimental psychologist would ever pose a question about comparing two such unrelated probabilities as in the example given above? In any event, the only sorts of probability comparisons that enter into anyone's axioms and theorems are those which Keynes accepted in one of his three categories.

Keynes' axiom system can be faulted on other, noncommon grounds, among them that it contained an uneconomically and inelegantly long list of definitions and axioms before he set about deriving theorems. The most interesting, and troublesome, feature of his definitions and axioms, however, is that they make the resulting system coincide with the classical probability calculus. One problem lies with the result, the other with his justification. Given the thoroughly epistemic grounding of his theory, a nonadditive calculus would have been more natural for many of the contexts he contemplated. He struggled with the idea of defining the lower endpoint of the scale as impossibility, noting that it pertained to fact rather than to knowledge, and Bateman (1987) observes that Keynes derived a formula for the probability of a generalization which adumbrates Cohen's inductive probability; but a scale with 0 representing ignorance or lack of evidence was evidently still too radical to have presented itself.

Keynes was inclined to doubt, in fact, that much could be accomplished in the way of assigning numbers to nonfrequency probabilities. In the *Treatise on Probability*, he considered carefully what was to become the standard proposal of measuring probabilities by betting quotients, but he took a pessimistic view of the procedure as an epistemological tool. Later (Keynes, 1933), in a review of Frank Ramsey's work, he reconsidered his old position on betting quotients but was never persuaded that subjective probabilities were rational.

There is one further criticism of Keynes' theory which is worth considering at this point, for its subtlety and fundamentality. It has been made by Ayer (1957/1962), among others. He poses the following problem: If probability is an objective relation between a conclusion and a particular body of evidence, then suppose we have several different sets of data with which to evaluate the same conclusion. There will be objective probabilities assignable to the relation between the conclusion and each set of supporting data, perhaps all of them different. Given that all these probability estimates are valid, what possible basis could we have for preferring one over another? In particular, how could we argue for undertaking any effort to increase the

evidence bearing on the conclusion? Our new estimate will not only be no more valid; it will also not even necessarily increase the probability of the conclusion. The new data might be unfavorable, or, no matter how relevant, they might leave us still as undecided as before.

Keynes himself provided the key concept for an answer to Ayer's objection, though he remained curiously undecided about its importance. He distinguished between the probability of a proposition, as the degree to which it is supported or weakened by given evidence, and the *weight* of an argument, as the total amount of available evidence⁶: "The weight, to speak metaphorically, measures the *sum* of the favourable and unfavourable evidence, the probability measures the *difference*" (p. 77). Other things being equal, it is always to our advantage to increase the weight of evidence bearing on a proposition we are undecided about; but, as Keynes clearly recognized, an increase in weight may be accompanied by a change in probability in either direction or by no change at all. (His own doubts about the concept pertained primarily to the *cost* of securing additional evidence, which necessitates the unrealistic qualifier "other things being equal"; clearly in most cases we reach sooner or later a point of diminishing returns in the collection of additional data.) All the estimates we have may be valid in their own right, but it is on account of the greater weight (not probability) that we properly favor the assessment based on the widest field of evidence available. The validity of estimates made on a more restricted basis rests on their contexts exhausting the field of evidence available, say, to a given subject.

The answer to Ayer's objection has also been nicely elaborated by Errol Harris (1970), without special reference to Keynes' concept of weight. He rests his answer on a view of knowledge as a highly integrated network of mutual implications. Ayer's puzzle arises for him only if propositions and their supportive evidence are regarded as isolable items of information which bear only an "external" relation to each other (such as a certain observed frequency of conjunction). In fact, the different bodies of evidence to which Ayer refers are not ordinarily independent. Even the very meaning of a proposition changes in the light of new information, so that what it is that is said to be more probable under new evidence is really a richer content of knowledge.

In the light of this fuller insight, the first proposition is less probable, but it is not wholly false, for it has been preserved and developed in the second, which is therefore preferable because more complete. The second hypothesis is not just an independent alternative to the earlier one, but is a better version of the same body of fact. (1970, p. 346)

Harris was not defending Keynes in particular, and in fact no one ever has. Primarily because his treatment was so reasoned and cautious, Keynes did not get so far as less circumspect thinkers. Part of the reason for his neglect is also that he

⁶Interestingly, few probability theorists before Keynes or since have used the concept of weight. Keynes (1921/1973) mentions Meinong and A. Nitsche as having made incidental use of such a concept, but Bernoulli (1713) and Peirce (1878b) also used the concept of weight of evidence, although Keynes overlooked these particular references. More recently, a similar concept has been used by Shafer (1976) and Jonathan Cohen (1977).

was writing in advance of the work of Fisher and others in statistics, and he did not make it easy for anyone of more reckless ambition than he to build on his very limited base.

It is interesting to note, finally, with Gillies (1988), that, just as Laplace did not use the Rule of Succession in his astronomical work, so Keynes did not use his probability theory in his work in economics. He was specifically skeptical, in fact, of econometrics and, according to Bateman (1987), wrote a trenchant critique, far ahead of its time in 1938, of the use of multiple regression in economics, citing problems of multicollinearity, serial correlation, linearity, measurement, and the absence of relevant frequency distributions and of long-term stability, among others (see Chap. 9).

One has to be constantly on guard against treating the material as constant and homogeneous. It is as though the fall of the apple to the ground depended on the apple's motive, on whether it is worthwhile falling to the ground, and whether the ground wanted the apple to fall, and on mistaken calculations on the part of the apple as to how far it was from the centre of the earth. (quoted in Bateman, 1987, p. 112)

8.1.2 Jeffreys

Jeffreys (1939/1961) is among those who criticize Keynes for “an unwillingness to generalize the axioms [which] prevented [him] from obtaining many important results” (p. 25n). Having started from an initial conception of probability quite similar to Keynes’, he went on to develop his theory much farther in the direction of mathematical statistics for scientific application.

He began, like Keynes, by embedding the subject of probability in the concerns of induction and logic generally. He was careful not to claim, either as his goal or his achievement, the *justification* of induction: “I do not consider justification necessary or possible; what the theory does is provide rules for consistency” (p. 424).

He lay down at the outset eight rules of theory construction (consistency, parsimony, etc.), of which the following are of particular interest:

3. Any rule given must be applicable in practice. A definition is useless unless the thing defined can be recognized in terms of the definition when it occurs. The existence of a thing or the estimate of a quantity must not involve an impossible experiment.

4. The theory must provide explicitly for the possibility that inferences made by it may turn out to be wrong. A law may contain adjustable parameters, which may be wrongly estimated, or the law itself may be afterwards found to need modification. It is a fact that revision of scientific laws has often been found necessary in order to take account of new information—the relativity and quantum theories providing conspicuous instances—and there is no conclusive reason to suppose that any of our present laws are final. But we do accept inductive inference in some sense; we have a certain amount of confidence that it will be right in any particular case, though this confidence does not amount to logical certainty. (1939/1961, pp. 8–9)

The intuitionist rule 3 is interesting because it specifically excludes all definitions of probability in terms of infinite sets, such as the Venn limit or the Mises collective. Rule 4 is “the chief constructive rule”:

It declares that there is a valid primitive idea expressing the degree of confidence that we may reasonably have in a proposition, even though we may not be able to give either a deductive proof or a disproof of it. In extreme cases it may be a mere statement of ignorance. (p. 15)

The reasonable degree of confidence is Jeffreys’ fundamental notion and his concept of probability. For him, as for Kahle and Keynes, probability is relative to evidence. He departs fundamentally from Keynes, however, in insisting that all probabilities are comparable, essentially that they exist on a single continuum. His first axiom is thus the following: “Given p , q is either more, equally, or less probable than r , and no two of these alternatives can be true” (p. 16; original in italics). Notice that he has restricted possible comparisons to those based on the same data p ; he contends, plausibly, that the practical cases where we need comparisons on different data comprise a rather special set (Keynes’ examples being mostly academic) and can be accommodated by special theorems (e.g., Bayes’ Theorem). There are six further axioms, of which the third identifies the extreme degrees of probability with certainty and impossibility.

Up to this point, numbers have not been used, and Jeffreys thinks their introduction not strictly necessary to further development of the theory, but mathematical expression is overwhelmingly *convenient*. He then introduces numbers by means of three *conventions*.

Convention 1. We assign the larger number on given data to the more probable proposition (and therefore equal numbers to equally probable propositions).

Convention 2. If, given p , q and q' are exclusive, then the number assigned on data p to “ q or q' ” is the sum of those assigned to q and to q' . (p. 19; original in italics)

Convention 3. If p entails q , then $P(q|p) = 1$ (p. 21; original in italics).

If we were to use, say, e^x in place of x as the probability of proposition p , then the addition rule (Convention 2) would be changed to a product rule; the probability of p or p' would become $e^{x+x'}$ instead of $x + x'$. A monotonic increasing function is required by Convention 1 and Axiom 1, each such function (like e^x) leading to a different rule for expressing the probability of a disjunction. But the same results would follow under any system. Abandoning Convention 1 would likewise make no material difference, but merely arrange all probabilities in the opposite order. The particular conventions Jeffreys selected are of course the most convenient and facilitate a match with the traditional calculus. The Laplacean measure of probability follows from his axioms and conventions for numerical measurement: “Given that a set of alternatives are equally probable, exclusive, and exhaustive, the probability that some one of any subset is true is the ratio of the number in that subset to the whole number of possible cases” (p. 23; original in italics). The association of probability with numbers does not, however, affect its definition.

The probability, strictly, is the reasonable degree of confidence and is not identical with the number used to express it. The relation is that between Mr. Smith and his name “Mr. Smith.” A sentence containing the words “Mr. Smith” may correspond to, and identify, a fact about Mr. Smith. But Mr. Smith himself does not occur in the sentence. (p. 20)

A number of writers (e.g., Neyman) have criticized Jeffreys’ theory on the grounds that if relative frequencies are to come out of a probability calculation, only relative frequencies can be put in. Jeffreys would be the first to agree: We *want* to get something out of a probability calculation besides relative frequencies; these are useless unless they are taken as a measure of support for a proposition. Jeffreys could turn the charge around by asserting that we cannot get measures of support out of probability calculations unless we put them in at the beginning.

The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency. . . . In many cases the numerical assessment [of a reasonable degree of belief] is the same as that of a corresponding frequency, but that does not say that the probability and the frequency are the same thing even in these cases. The fact that physicists describe an atmospheric pressure as 759 millimetres does not make a pressure into a length. . . . The notion of a reasonable degree of belief must be brought in before we can speak of probability; and even those writers that do not mention it at the beginning have to use it at the end before any application can be made of the results—or else avoid the question by allowing the person advised to supply it himself, which he does in practice without the slightest difficulty. Even if the prior probability is *based on* a known frequency, as it is in some cases, reasonable degree of belief is needed before any use can be made of it. It is not *identical with* the frequency. (pp. 401–402)

There is one incidental passage in Jeffreys’ book which leaves him open to a charge of inconsistency in application of his definition of probability. In this passage, where he remarks that it would be incorrect to speak of the probability that the 10,000th decimal of e is a 5 as .1, he takes instead the position of the extreme frequentists, such as Neyman:

By following the rules of pure mathematics we could determine it definitely, and the statement is either entailed by the rules or contradicted; in probability language, on the data of pure mathematics it is either a certainty or an impossibility. (p. 38)

In a footnote to this passage, he goes on to criticize pure mathematicians for speaking of a probability distribution of prime numbers. The objection is sensible (and very commonly pointed out) from a frequency view, since there is no long sequence of conceivable trials; but it is difficult to understand what the problem is for Jeffreys. If I have no idea what the 10,000th decimal of e is (as is the case), then surely a uniform distribution over the 10 possible values represents as reasonable an allocation of prior probability as it would in the case of any other totally unknown decimal quantity. Jeffreys is correct—and consistent with his own theory—in saying that it is either a certainty or an impossibility that the 10,000th decimal of e is a 5 “*on the data of pure mathematics*”; the problem is that if we are allowed a similar statement of conditional data, nearly everything in the universe is a certainty or an impossibility. On some appropriately selected data, it is either a certainty or an impossibility that the mean of a given population is zero; and so on, to all the other

examples to which he applies his probability calculus. Jeffreys can reply that I do not have the option of making a “raw” probability assessment of the 10,000th decimal of e without taking the data of mathematics as given, because they are inescapable: My existing knowledge still *entails* a particular value, even if I do not at this moment know what it is. But again, my existing knowledge, if I followed through all its implications, may well be sufficient to entail a particular value for the mean of the population in question. The only way to keep this objection from hounding Jeffreys’ every application, and nullifying his entire theory, is evidently to allow a radical contingency into the empirical world which is debarred from mathematics. If this is what Jeffreys is thinking, he does not say so. In fact, his attempt at clarification only deepens the confusion, for he says that “probability is not a guess”; but it is difficult to imagine what else, of greater epistemological glory, probability could be in situations—routine in his own applications of the theory—where we have no prior knowledge at all and call in the principle of indifference to make the assignment.

This objection, whatever its merit, cannot be the reason for neglect of Jeffreys’ theory, because it does not seem to have been generally noticed. Popper (1968, p. 375n), at the end of a footnote in appendix *vii, criticizes Jeffreys’ attack on the idea of a probability distribution of prime numbers; he says that if we put balls into an urn which are numbered from 1 to n , and consider the problem of the probability of drawing out a ball with a prime number, any reasonable theory of probability will have to give the answer that the probability in question approaches zero as n becomes large. But he does not advert to the fact that Jeffreys, of all probability theorists, is an odd one to reject the problem as meaningless, nor does he pursue the implication that application of Jeffreys’ theory must be limited to events of objective, or metaphysical, chance.

The defect is not a fatal one to the theory; it could be remedied easily enough by accepting these two problems as meaningful within his theory; but Jeffreys did not acknowledge the inconsistency which would make the remedy necessary.

Altogether, Jeffreys’ contribution was enormous, however its individual aspects may be appraised. He provided Bayesian counterparts to most of the techniques taught in an introductory statistics course and to a good many others. At the same time he gave careful attention both to foundations and applications. He aimed at an explicit axiomatic development of the theory, but, as a physical scientist rather than a pure mathematician or philosopher, he always had the needs of the research worker in mind; and he aimed for his theory to model ideally the mental processes of scientists learning from experience, showing them how to use new findings to modify their previous opinions. Hacking (1965) says unequivocally that:

Jeffreys’ *Theory of Probability* is the most profound treatise on statistics from a logical point of view which has been published in this century. It marks the first attempt to apply the axiomatic procedures, familiar to logicians, in deriving, criticizing, and contributing to the methods of statisticians. (p. 201)

8.1.3 Jaynes

E. T. Jaynes (2003) updates Jeffreys with a recognizably modern twist. A major problem with Jeffreys' logical theory of probability, in the eyes of most appraisers, was who would make the logical probability judgments. Jaynes aimed to develop a theory of inductive inference for a robot. In the process, he saw himself also a providing a proof that ideal reasoning was Bayesian, and his purported proof seems widely to have been accepted as successful.

Jaynes brings a keen mind to the task. One impressive piece of evidence is his quick dispatch of Bertrand's paradox (Chap. 6). Jaynes proceeds by identifying three assumptions which have remained implicit in all previous discussions of Bertrand's problem, but to which any reasonable person would readily assent: that the solution should not depend on the position from which the circle was viewed or in the size or location of the circle. It turns out that Solution 2 is uniquely compatible with translational invariance and is also compatible with rotational and scale invariance. It is remarkable that Jaynes was the first to make this discovery in a century, and it appears to be typical of his skills in solving practical problems. Jaynes, like Jeffreys, was a physicist and engineer.

Despite his impressive analytical skills, Jaynes' program of developing the structure of inductive reasoning for a robot fails, for rather predictable reasons. In the first place, to the robot, as a nonliving entity, nothing matters; nothing has any meaning. Jaynes readily acknowledges that he supplies the meanings. The robot doesn't do semantics (p. 94 n). Although nothing has any meaning to the robot, Jaynes wants its input restricted to propositions with unambiguous meaning (to him, presumably)—no terms with disputed meanings, like *speech*, or *militia*, or *involuntary servitude*—or *probability*. He also excludes conflicting evidence—to spare the robot the “agony” (p. 18) of reasoning from inconsistent data. No ambiguities, no conflicts—how much of the universe is left? Jaynes thinks of his robot as virtually human, but not an autonomous one. He wants the robot to behave like child, or a slave, and “believe” everything he tells it; a skeptical robot would be “dangerous” (p. 99). It is essentially a calculating machine (for what it accomplished, one would think pencil and paper would do). The robot will take account of all the (unambiguous, consistent) evidence, he says, and be purely unbiased in its conclusions. All the evidence, that is to say, as we present it. The most telling example is Jaynes' discussion of Soal and Bateman's ESP research. Results with one of their participants had a *p* value of 10^{-137} ; like C. E. M. Hansel (1966), Jaynes takes this as proof of cheating. He says straightforwardly that he would choose whatever prior probability was necessary to swamp that 10^{-137} ; nothing could convince him that ESP was real. If that isn't sheer prejudice, I don't know what is. Perhaps Jaynes would defend himself by acknowledging that the prejudice resided in him rather than in the robot, but then it's not clear what he needs the robot for—except to make his prejudice look more scientific. Jaynes assures us that evolution selects for Bayesians, but I think his robot, if it were alive, wouldn't last long.

Jaynes sees himself as having provided the inescapable, unique, correct rules for uncertain inference—or, to use the term he borrowed from Polya (1954)—plausible reasoning. And his results seem to be generally accepted as such, though not all the work is original with Jaynes; in particular, he borrowed from Cox (1961). Both start from the proposition that the probability of (A or \bar{A}) must be 1; they build in as a fundamental assumption that a probability of 1 must represent certainty and a probability of 0 must represent impossibility, ensuring the familiar additive frequency scale. But in doing so, they thoroughly confound the *epistemic* with the *alethic*. Although it is true that either a proposition or its denial must be *true*—excluding those with imaginary truth values (Spencer Brown, 1972), like “This sentence is false”—our *knowledge* about either A or \bar{A} , or both, may be 0. Using the language of personalist Bayesians (cf. *infra*), we could say that Jaynes’ system—Bayesian theory in general—forces us to commit all our chips on every proposition, since knowledge, in this vision, is a matter of betting against a malevolent Nature, seeking to win against us, and we are forced to pay for incorrect guesses. If we don’t stake the full quota of belief every time, we are guaranteed a long-run loss. That would be irrational: Q.E.D. But who ever said we had to play their stupid game? If the Scientific Service System disallowed conscientious objector status, then Bayesianism would be a rational betting strategy; but it’s a separate question whether such a betting system has anything do with the state of our knowledge. Jaynes is treating his calculus of truth as a calculus of support. His calculus has no way of expressing ignorance, or lack of support, except by committing an equal amount of support, or belief, to all the alternatives.

Jaynes is aware of the obvious problem here, which can be framed as a matter of the weight of evidence. His own example is as useful as any. Suppose, on the one hand, that he has a coin which he has physically examined and determined to be balanced. Then his assignment of a probability of $\frac{1}{2}$ for a head is based on a large weight of evenly balanced evidence. Now suppose the question is about the probability of life on Mars. In this case he has (evidently) no evidence at all, and he also assigns a probability of $\frac{1}{2}$. But to say that life on Mars is as likely as not to exist as not to exist is not to make a statement of ignorance, but a highly specific claim. (I think Jaynes’ example predated the mission to Mars, but his probability of $\frac{1}{2}$ still surprises me. But the particular value doesn’t matter, and the value of $\frac{1}{2}$ makes a handier example.) Jaynes proposes to handle the situation by assigning a distribution, A_p , to the probability. In both cases the mean would be $\frac{1}{2}$. In the case of the coin, the distribution would be sharply peaked around $\frac{1}{2}$, perhaps like a beta distribution with $r = 50$ and $N = 100$. For the question of life on Mars, his A_p would be U-shaped, like a beta with $r = N = 0$. The weight of evidence is then reflected (inversely) in the variance of the A_p distribution and will correspondingly be taken into account in the posterior distribution which incorporates the imaginary N generating the A_p distribution. Jaynes is uncomfortable with the idea of a probability of a probability and tries to deal with it by creating an “inner” and “outer” robot, corresponding, he says, to the subconscious and conscious mind, so that the two probabilities are on different levels. He is fierce throughout the book in denouncing the

“adhockeries” of frequentism, but this seems an unfortunate piece of adhockery itself, at the foundations.

The metaphysical status of probability is an issue throughout Jaynes’ system. As for Bernoulli and Laplace, probability is for Jaynes a matter of limitations in our knowledge. Randomness, in particular, is just an acknowledgment that something is more complicated than we want to bother with. On this view, any sort of randomization, of *introducing* randomness, of deliberately creating uncertainty, is ridiculous and reprehensible. The idea of randomness as an attribute of nature he regards as the “mind projection fallacy.” However, we have already seen that Jaynes thoroughly confounds matters of objectivity and subjectivity at the foundations of his theory, when he takes rules of logic to be rules of thought. We might call this the *world projection fallacy*.

Jaynes’ robot, reasoning according to fixed rules, as a means of replacing the personal with the logical, appears as the ultimate realization of the seventeenth-century dream of depersonalizing knowledge. And the closer we get to that ideal, the sillier it looks. Does it make any more sense to take Bayesian rules of inference as a model of thought, ideal or actual, than to say that differential equations represent our thinking about fluid dynamics? Or that the operations of long division represent our thinking about partitioning? And, if our goal is either to understand human reasoning or to model it, how much sense does it make to start by eliminating the *person*?

Whatever faults we find with Jaynes’ theory, we can at least give him credit, like Ramsey, for fulfilling the wish of Kendall (1968) that all Bayesians follow the model of Bayes himself by publishing posthumously. Jaynes died in 1998; his book was prepared for publication, conservatively, by his student Larry Bretthorst.

8.2 Personalist Theories of Probability

The logical view of probability can be seen as leading almost ineluctably to a subjective or personalist view. The problem on the logical view is just who is to say which degrees of belief are reasonable. Jeffreys supposedly once said that probabilities should be laid down by an international body; Good (1976) has added that Jeffreys would undoubtedly be the chair. Criteria like coherence may be, and have been, advanced, but the criteria themselves are then in need of judging. The subjectivists have not shrunk from the apparent implications; for them, probability is *Your* degree of belief. (Use of the capitalized second-person pronoun is the terminology of de Finetti, 1974.) The general subjectivist program is thus to formalize the judgments of individuals, but—lingering shades of the logical theory—by means of imposing certain conditions (e.g., consistency) to make them somewhat idealized probability appraisers.

The subjective branch of probability theory has some long roots. In the days when the concept of mathematical probability itself was just in process of formation, and the principal applications were to gambling and annuities, problems were

more commonly posed in terms of expectation than of probability (Chap. 3). Although from an abstract point of view, the difference is immaterial—either concept can be defined in terms of the other—the trouble with making expectation the primitive concept is that we are not always indifferent to alternatives with the same expectation. Counterexamples are provided by the bird-in-the-hand principle: We may well prefer a sure gain of \$1000 to a 1000-to-1 shot at \$1000,000, though the expectations are the same. The Petersburg problem provided a particularly compelling example.

For this reason expectation was wholly displaced by probability as the fundamental concept over two centuries ago. Interest in the expectation approach was revived, independently and at about the same time, by several different thinkers, including Émile Borel in France, Frank Ramsey in England, and Bruno de Finetti in Italy, in connection with subjective theories of probability. Popularization of the subjectivist approach in statistics has been more the work of I. J. Good in England and Leonard Savage in the United States. In view of the differences among all of these theorists, objection is sometimes raised to speaking of *the* subjectivist theory. Good's classification of 46,656 types of Bayesian theory gets a little silly, since there aren't that many Bayesians in the universe, but we might agree at least with Pratt's (1965) gross distinction between Good Bayesians and Savage Bayesians.^{7,8} I shall concentrate here on the work of de Finetti, since he has developed the theory most fully; but his influence was delayed by the fact that his earliest papers, from the 1920s, were published in Italian, and his major paper (de Finetti, 1937), in French, was not translated into English until 1964 (in Kyburg & Smokler, 1964). Just by virtue of his being English, Ramsey was more influential than the size of his corpus⁹ would warrant, and we shall first have a quick look at his ideas.

8.2.1 Ramsey

Ramsey, a brilliant student of Russell and Wittgenstein, could fairly be called the Galois of probability theory: He died in 1930, of jaundice,¹⁰ at the age of 26. His work was published posthumously as a collection of essays entitled *The Foundations of Mathematics* (1931). Ramsey, like Borel, got started on subjectivism from

⁷Good (1976) erroneously attributes the terms to Maurice Bartlett.

⁸Others contend that there is no problem in classification since different kinds of Bayesian can be readily distinguished by their posteriors (McGrayne, 2011, p. 129).

⁹14 stone

¹⁰There has always been a lot of speculation about the cause of Ramsey's death. The jaundice, observed in November 1929, was obvious, but the cause was not. Misak (2020), reviewing the evidence, thinks it was Weil's disease, acquired from swimming in the River Cam, which Ramsey loved to do. Normally the river would have been too cold in October either for swimming or for keeping the bacteria alive, but the fall of 1929 was unusually warm and sunny (and for that reason one of the best years of the century for wine, all across Europe).

critically contemplating Keynes' theory; Keynes himself, as was noted earlier, appears ultimately to have been persuaded that Ramsey was on the right track. Ramsey, for his part, questioned whether there is, in fact, any such thing as Keynes' "probability relations":

All we appear to know about them are certain general propositions, the laws of addition and multiplication; it is as if everyone knew the laws of geometry but no one could tell whether any given object were round or square; and I find it hard to imagine how so large a body of general knowledge can be combined with so slender a stock of particular facts. (1931, p. 162)

He went on to say that "No one estimating a degree of probability simply contemplates the two propositions supposed to be related by it; he always considers *inter alia* his own actual or hypothetical degree of belief" (p. 163). He took up straight-away the question of how degrees of belief, as psychological states, may be measured, and decided to limit himself, for reasons of expediency, to beliefs which have some relevance for *action*, thus giving him a behavioral handle on degrees of belief. His basic premise was then that:

We act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions. This theory cannot be made adequate to all the facts, but it seems to me a useful approximation to the truth particularly in the case of our self-conscious or professional life, and it is presupposed in a great deal of our thought. (p. 173)

Expectation thus becomes the central concept—we act on the basis of maximum expectation, and degrees of belief can be derived as the coefficients by which the "goods" are multiplied to yield the expectation. Ramsey suggested that betting provides a paradigm for the evaluation of expectations themselves; we can tell people's expectations from their betting behavior, and their state of belief from the odds they will just take on some proposition, in view of some proffered good contingent on its truth.

Ramsey was aware of several problems with his approach. For example, we must take account of a person's bet being influenced by *wishing* a proposition to be true; he proposed that we must start with an "ethically neutral" proposition as a reference point and stipulated as an axiom that some such proposition exists for every individual. (Bayesian theorists have since found other ways around the problem.) Again, problems of noncomparability arise if we place no restrictions on betting. A system of bets may be placed on a set of alternatives such that the total probability is less than 1. This contingency is covered by the rule against "Dutch book": An individual is not allowed to place bets in such a way as to assure a net loss, no matter what happens. (The Dutch book condition has been criticized by Ellis, 1973, who is himself sympathetic to Bayesianism, but it is still the standard solution.) Other problems are essentially bracketed. Ramsey simply assumed an individual to be indifferent between expectations having the same value. He limited himself to a consideration of finite alternatives. The connection between his calculus of degrees of belief and the traditional calculus of frequencies was accomplished in a sentence: "Supposing goods to be additive, belief of degree $m/n \dots$ is the kind of belief most appropriate to a number of hypothetical occasions otherwise identical in a

proportion m/n of which the proposition in question is true" (p. 188). Similarly with statistical applications: "(For all this, see Fisher)" (p. 204).

Ramsey did not accomplish much beyond spadework. His contribution was principally to have been the first modern thinker to take the idea of subjective probability seriously. He never had time to follow it through, and his work was published in unfinished form.

8.2.2 *De Finetti*

The personalization of probability was carried through most thoroughly by de Finetti, who saw probability as arising from uncertainty, whatever its source. Hence a string of digits in the decimal expansion of π —say, the 2001st to the 3000th—qualifies, as well as a string of digits in a random number table, as an object of probability theory. Indeed “random,” in his theory, means “unknown to You,” regardless of whether the event is determined or known to anyone else. It is a consequence of de Finetti’s definition that a probability cannot be unknown: Probability characterizes uncertainty, and there is no second-order uncertainty about uncertainty. He compares the concept of unknown probability with “thinking that in a statistical survey it makes sense to indicate, in addition to those whose sex is unknown, those for whom one does not even know ‘whether the sex is unknown or not’” (de Finetti, 1974, p. 84).

In using betting to measure probabilities, de Finetti is of course involved in the comparisons of expectations. Indeed, like Borel (1924), he introduces probability and expectation through the explicitly monetary concept of prices. If X is a “random gain,”

We might ask an individual, e.g. You, to specify the *certain gain* which is considered *equivalent* to X . This we might call the *price* (for You) of X (we denote it by $P(X)$) in the sense that, on your scale of preference, the random gain X is, or is not, preferred to a certain gain x according as x is less than or greater than $P(x)$. (1974, p. 73)

He stipulates various properties, such as additivity, for the price operator P ; then, preserving the same notation, he defines probability analogously to a price: “The probability $P(E)$ that You attribute to an event E is therefore the certain gain p which You judge equivalent to a unit gain conditional on the occurrence of E ” (p. 75). In comparing expectations, de Finetti faces the same problem as Ramsey and his predecessors; he meets the problem simply by stipulating rigidity for the scale of utility over the range of values likely to be encountered. This assumption is necessary to insure that “You should be indifferent between ‘receiving with certainty a sum S , or twice the sum if a particular one of the two possible cases occurs’” (p. 78). De Finetti compares this condition to the assumption of rigidity for solid bodies over normal temperature ranges; the theory can presumably be modified ad hoc to handle peculiar cases like the Petersburg problem.

Another aspect of the determination of preferences is the condition of gain or loss. De Finetti defines expectation and probability with respect to a quadratic loss function: If X is a random quantity, then Your expected value, \bar{X} , of X is that value You would choose if the loss You suffered were proportional to $(X - \bar{X})^2$. We want loss to be positive for errors in either direction; to minimize loss analytically, we want to be able to differentiate the loss function; but the absolute value function has a corner and is not differentiable, whereas the square of the loss is twice differentiable. Quadratic loss as a criterion of coherence also has a convenience similar to that enjoyed by the variance in normal distribution theory; if we set up a space of expectation points and outcome points, then the loss for a given bet is proportional to the squared (Euclidean) distance between the respective expectation point P and outcome point Q . To make Your bets coherent, You must choose P so as to minimize the distance to all points Q . De Finetti offers two reasons for favoring the criterion of minimization of loss rather than maximization of gain in the determination of probabilities. The first is historical: The definition of loss is due to Wald, who expanded Neyman-Pearson theory to include economic decision-making (see below). The second is strategic: He figures that individuals will be more cautious in their assessments if they can only lose than if they can only gain.

It is a remarkable feature of subjective theories of probability that probability assessments invariably turn out to correspond to the Laplacean probabilities of the traditional calculus. We should notice how this happens in de Finetti's theory. In the situations considered by classical probability (dice, coins, etc.), de Finetti agrees that the essential feature is symmetry of alternatives, which leads to an equal partitioning of probability in line with that achieved by the principle of indifference. The difference of de Finetti's theory is principally to insist on a subjective slant: Whereas the classical theorists attempted—vainly—to ground the principle in some objective identity or symmetry of alternatives, he takes a constructivist stance, contending that identity is a matter of the individual appraiser's grouping. Similar things happen with respect to independence. Instead of postulating a physical independence between events like coin tosses, he adduces the concept of *exchangeability*, which is essentially invariance of the probability of a sequence of events with respect to their order. Though exchangeability is a somewhat weaker condition than classical independence, de Finetti is able to prove the standard theorems using it instead, and therewith makes incidentally a contribution of some value to probability theory at large.

In a theory of *personal* probability, the conformity of probability to relative frequency raises interesting questions about the status of probability. Surely it is not, after all, *Your* beliefs—or anybody else's—which are being formalized by this theory. Ramsey (1931) noticed this much. Chances, he said, “do not correspond to anyone's actual degrees of belief; the chances of 1,000 heads, and of 999 heads followed by a tail, are equal, but everyone expects the former more than the latter” (p. 206).

De Finetti, on the other hand, discussing the example of decimal digits being selected at random, says,

Even after 100 or 1000,000 or 10^{1000} consecutive zeroes, provided we have no gift of divination, the probability that the next figure will be zero is $1/10$, as for any other figure; the probability that the next 100 figures will all be zero is 10^{-100} , as for any other 100-figure number; the probability that the figures will continue to be zero for evermore is zero, exactly as it is at any other instant, and after any arbitrary sequence of figures. (1974, p. 127)

But, as Ramsey was aware, in point of fact we should, in actual such circumstances, begin to modify our expectation of a zero before even the 20th consecutive zero. The way de Finetti speaks of “*the probability*” makes it sound much more objective than subjective, and so it is hard to construe his concept of probability as strictly personal. Elsewhere he describes his theory of probability as “*a normative theory for coherent behavior*” (p. 192n; his emphasis) and specifically rejects as irrelevant any criticism that his theory is inconsistent with actual human reasoning; he argues that such “mistakes” in probability appraisals are no more relevant to the theory of probability than arithmetic errors are to the science of mathematics. In this respect, de Finetti’s view is exactly like Jeffreys’, except that Jeffreys defines his probabilities in the first place as logical rather than personal. Matalon (1966) virtually accuses the personalists of maintaining a double standard: They define probability relative to individuals’ actual beliefs; but when confronted with examples where the beliefs of apparently reasonable individuals conflict with the theory (he refers to Allais, 1953; see below), they retreat behind the freedom of the mathematician to construct arbitrary axiomatic systems. It is hard to know what happened to You with a capital Y.

Consider two more passages bearing on the issue. As if to clarify his own phrasing in the preceding quotation, de Finetti argues (quoting Jeffreys) that realistic, objective modes of expression are simply what we are forced into by natural language and that we should not be misled by this property of language into a realist way of thinking.

It is somewhat natural, when stating a problem, to use expressions like: “Suppose that the probability of E is p ”; “Suppose that these events are independent”; and so on. But such expressions can properly refer to nothing other than the opinion of the individual concerned, whether myself, or somebody else, or an imaginary person maintaining some opinions no matter how reasonable or foolish they are to us; and what the expressions convey—in a slightly objectionable form—is simply that the probabilities and properties mentioned are supposed to describe the opinion of that individual. (de Finetti, 1972, p. 192)

But somehow, at the same time that probabilities express degrees of uncertainty for You,

The calculus of probability can say absolutely nothing about reality; in the same way as reality, and all sciences concerned with it, can say nothing about the calculus of probability. The latter is valid whatever use one makes of it, no matter how, no matter where. One can express in terms of it any opinion whatsoever, no matter how “reasonable” or otherwise, and the consequences will be reasonable, or not, for me, for You, or anyone, according to the reasonableness of the original opinions of the individual using the calculus. As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before (either in terms of having accepted certain facts, or having evaluated degrees of belief in them, respectively). (1974, p. 215; his emphasis)

Perhaps what de Finetti intends is to distinguish sharply between the *calculus* of probabilities as a set of nonsubstantive rules, and empirical probability *assessments* by individuals. Such an interpretation would appear to be consistent with the meaning of the preceding passages. But it is hard to see how such a sharp line can be drawn. The rules set clear constraints on what particular probability assessments may be; it is far from being the case that just any set of beliefs can be plugged into the calculus. Still, this distinction may hold the key to understanding the logical/empirical status of probability in this theory. The *meaning* of probability is Your uncertainty with respect to any unknown, as measured by the bet You would just accept under conditions of quadratic loss, etc. The supposedly nonempirical part of the theory is the set of further conditions—such as conformance with classical probabilities—which You must meet if You would be coherent in Your evaluations. De Finetti can speak impersonally of *the* probability of an event only as misleading ellipsis in a realist language, about situations of nearly universal consensuality. In other places, when he is emphasizing the personal nature of probability, he cautions that this feature must not be taken as relieving the probability appraiser of any responsibility for the most considered judgment, that “subjective” is not to be understood as “arbitrary.” The formal theory cannot be counted on to correct poor judgments; what it does purport to do is to provide a standard of coherence for even responsibly assessed initial probabilities.

8.2.3 Savage

It will be useful to compare the views of Leonard Savage (1954) on the logical status of personal probability. Savage described his own theory as “a highly idealized theory of the behavior of a ‘rational’ person with respect to decisions” (p. 7). Part of rationality, in deductive situations, is conformity with the rules of logic; and part, suggested Savage, is conformity with his theory of probability. He invited us to consider his theory as a possible model of rational decision-making in situations of uncertainty. Logic, like probability, he noted, can be taken in both a normative and an empirical sense, as a statement of how people ought to behave and as a crude and approximate description of their actual behavior. The two often diverge, however; and for Savage, as for de Finetti, the relevant aspect was normative.

According to the personalistic view, the role of the mathematical theory of probability is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it. An inconsistency is typically removable in many different ways, among which the theory gives no guidance for choosing. (Savage, 1954, p. 57)

Savage went on to consider an example: A coin is to be tossed five times. A person finds “on interrogating himself” about the 32 possible outcomes that he considers all of them equally likely. But he also regards it as more likely that a total of four or five heads will turn up than that the first two tosses will both be heads. And

“reference to the mathematical theory of probability” shows that the former event has probability 6/32, the latter 8/32; hence his personal probabilities are inconsistent. “The theory does not tell him how to resolve the inconsistency; there are literally an infinite number of possibilities among which he must choose” (p. 57). Savage’s solution was the observation that usually there are some probability relations we are more sure of than others, and we sacrifice the less sure to the more sure in the name of consistency. He resisted the temptation to formalize this step, as probabilities of probabilities, noting that it would involve us unprofitably in infinite regress: The resulting composite probabilities would presumably be subject in turn to varying degrees of sureness; and the whole theory becomes progressively unreal.¹¹

We see in this example that the “mathematical theory of probability” once again provides the standard followed by personal probability. Savage (1954, p. 33), in fact, took the usual axioms as starting points: A probability measure $P(B)$ is defined such that

1. $P(B) > 0$ for every B .
2. If $B \cap C = \emptyset$, $P(B \cup C) = P(B) + P(C)$.
3. $P(S) = 1$

At this point, Savage’s theory is distinguished from objective theories in the interpretation placed on the axioms and from other subjective theories with respect to several details. In the first place, he wanted to admit all probabilities as orderable, contra Keynes.

I personally consider it more probable that a Republican president will be elected in 1996 than that it will snow in Chicago sometimes in the month of May, 1994. But even this late spring snow seems to me more probable than that Adolf Hitler is still alive. (p. 27)

De Finetti would presumably agree with the orderability asserted here. But Savage parted from de Finetti in his handling of expectation and utility. Whereas de Finetti postulates linearity of utility over the necessary range, Savage solved the problem of variable utilities by the creation of a new unit, the *utile*, measuring equal increments of utility, whatever monetary (or other) value may be associated with them. This solution has its drawbacks, too, however; as de Finetti (1974) points out, if utility and money (or whatever) are no longer equated (except for a change of scale), then our measuring stick for financial transactions is constantly in flux. The cash value of a utile is continually dependent on the outcome of other transactions, and betting becomes somewhat chaotic.

De Finetti (1974) emphasizes the importance of acquiring a “feel” for degrees of subjective probability, and he recommends Good’s (1950) technique of imaginary experiments for this purpose. Savage, for his part, emphasized the need for acquiring a subjective sense of utility. He cited in this connection an embarrassing example due to Allais (1953). A person is offered two pairs of gambles:

¹¹ On similar grounds he declined to consider the cost of calculation in the evaluation of personal preferences: Is it worth my while in a given case even to think about undertaking complex calculations to determine my preference? and so on.

- A_1 : 1 million Dfl for sure;
- B_1 : 5 million Dfl with probability .1,
1 million Dfl with probability .89, or
nothing at all with probability .01;
- A_2 : 1 million Dfl with probability .11 or
nothing with probability .89;
- B_2 : 5 million Dfl with probability .1 or
nothing with probability .9.

Savage found, perhaps like many of us, that he preferred A_1 to B_1 and B_2 to A_2 ; but, as Allais pointed out, that combination is inconsistent with any expected utility maximization principle. Savage's response was essentially to contemplate the problem until persuading himself that he really did prefer gamble A_2 to B_2 after all, and his discussion seems to imply that we should be prepared to do the same in similar situations.

It seems to me that in reversing my preference . . . I have corrected an error. There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be. Let me illustrate by a simple example containing no reference to uncertainty. A man buying a car for \$2,134.56 is tempted to order it with a radio installed, which will bring the total price to \$2,228.41, feeling that the difference is trifling. But, when he reflects that, if he already had the car, he certainly would not spend \$93.85 for a radio for it, he realizes that he has made an error. (Savage, 1954, p. 103)

Savage's analysis is not unreasonable; what is unsatisfying about it is that, if personal probability allocations are always to be corrected by logical analysis, it is hard to see the force of the personalist definition; and Savage's theory verges on collapse into Jeffreys'.

8.2.4 Wald's Decision Theory

Before proceeding to a discussion of Bayesian statistical inference, it is appropriate to consider briefly the views of Abraham Wald. Strictly speaking, Wald does not belong either in this chapter or the preceding; he is included at this point for convenience because his work served as a bridge between the Neyman-Pearson and the Bayesian theories and helped substantially in the acceptance which Bayesianism has found in this country.

The stage for Wald's contribution was amply set by Neyman's own shift from inductive inference to inductive behavior as the object of his theory. In its original conception, the theory, involving the truth or falsity of hypotheses, had a distinctly realist flavor—even if truth was represented simply as a point in space. Within 10 years, however, the focus had shifted to “deciding to act *as if* the hypothesis were true,” with a more instrumentalist orientation. Gière (1976), in fact, attributes much of the success of the Neyman-Pearson approach to the preeminence of

instrumentalism in contemporaneous philosophy and science, and the same characteristic may have much to do with whatever acceptance is ultimately achieved by Bayesian methods.

Wald (1950) started from two principal dissatisfactions with the Neyman-Pearson theory: that as a decision theory it was limited to two possible decisions (accept or reject) as outcomes, and it was entirely concerned with single-shot experiments. His contribution was thus to generalize the Neyman-Pearson theory along several dimensions, toward a more abstract theory of “inductive behavior” (i.e., decisions): He broadened the decision space, introduced the concepts of risks and losses, and formulated the theory of sequential testing. The original theory of hypothesis testing thus remained as a special case; size and power were special instances of risk. Wald admitted that the concept of loss is a little vague when applied to decisions on scientific hypotheses; the cost of experimentation he handled simply by assuming it to be proportional to the number of observations taken.

The key point in Wald’s theory is the choice of criterion for decisions. Whereas Neyman and Pearson opted for a constant value for a particular type of risk, Wald, influenced by von Neumann’s work on the theory of games, gave almost exclusive consideration to the minimax rule (which Neyman & Pearson, 1933, also briefly considered). The minimax strategy—deciding in such a way as to minimize the maximum expected loss—is a conservative strategy which is regarded as theoretically ideal for two-person zero-sum games. When applied to the testing of scientific hypotheses, as Wald recognized, the implication is that we are playing against a malevolent Nature, who, using a minimax strategy herself, seeks to maximize our risk. Wald suggested, however, that such an assumption “is perhaps not unreasonable” (1950, p. 27); and, as a further possible defense, he proved (p. 125) that the minimax rule coincides with a Bayes solution—with a “least favorable” a priori distribution (i.e., the one for which the minimum risk is a maximum).

In considering the properties of Bayes solutions, Wald in no way wished to identify with the subjective theory of probability. He generally favored a mathematical approach, and scrupulously tried to stay above disputes over the meaning of probability. Nevertheless, all the special features of his theory—the explicit incorporation of risks, losses, and utilities, the commitment to a sequential program of decision making—led directly into the formulations being developed by the subjectivists. Modern Bayesian theory gained most of its impetus from applications to decision making in business; that was the subject of the first Bayesian text, by Schlaifer (1959), which, like Savage’s work, used a number of concepts originally put forth by Wald.

The evolution of Neyman-Pearson theory toward Bayesianism, via Wald, may have been a happy development for business,¹² but it has largely passed psychology

¹² Or maybe not. Kiefer (1977) notes that Bayesianism has been popular with American businessmen because they value decisiveness, and Bayesian procedures give them a rule for action; they never know how much they gain or lose by using this approach rather than some other. The advent of interactive computer packages, which, at least at least at the time Kiefer was writing, typically provided posterior distributions but not operating characteristics of decision strategies, insulated

by. It took several decades for textbooks in this field to absorb the first round of innovations by Neyman and Pearson in the Fisherian significance testing theory (Chap. 9). And the instrumentalist decision orientation, in so naked a form, seems bound to be more confusing than appealing, for some time to come. Occasional texts, like Hays (1963, 1981), have attempted to incorporate the new perspective, but the discussions of costs and utilities still often seem to be tacked on, rather than integral. A major reason, of course, is that the costs and utilities related to decisions in psychological research are more obscure than those in many business or everyday problems; and the situation is muddied still further by the fact that the decision orientation, however religiously espoused, was never taken very seriously in psychological research, and an epistemic attitude toward statistical analysis has prevailed.

8.3 General Structure of Bayesian Inference

One form of Bayes' Theorem in Chap. 4 was for the probability of a hypothesis H_0 , given some data D :

$$p(H_0|D) = \frac{p(D|h_0)p(H_0)}{\int p(D|H_i)p(H_i)}.$$

The denominator, ranging over all possible hypotheses (expressed in terms of the values of some parameters θ_i), amounts to a normalizing constant, so the formula may be rewritten:

$$p(H_0|D) \propto p(D|H_0)p(H_0).$$

The first term on the right, $p(D|H_0)$, is the only probability which appears in the frequentist theory: the probability of the data, given the hypothesis. In that theory, it is represented by significance levels and confidence coefficients. To keep the various kinds of probability in Bayesian theory distinct, it is known, within and without Bayesian theory, as the *likelihood* of the observations. This usage is different from that of Fisher, who introduced it into modern statistics to refer to the likelihood of a hypothesis or of some parameter value designated by the hypothesis. The second quantity on the right, $p(H_0)$, represents the unconditional probability of the hypothesis. It is known, for obvious reasons, as the *prior probability* of H_0 . In most

their calculations still further from reality. Kiefer's remarks are not altogether irrelevant to psychologists:

There are so many customers who do not understand fully the consequences of what they are doing in using such routines, to which they have been attracted by the oversimplified promise of such ease and comfort in reaching a quick decision to which such beautiful names are attached. (Kiefer, 1977, p. 171)

applications, it can only be a degrees-of-belief probability rather than a relative frequency. The resultant quantity, on the left, $p(H_0|D)$, represents the probability of the hypothesis conditioned on the data. This *posterior probability* of the hypothesis will be a Bayesian, degree-of-belief probability if the prior was.

Bayes' Theorem in this form exhibits neatly the essential structure of Bayesian inference, as a process of the progressive modification of opinion by data. If the hypotheses in question concern the value of some parameter θ , we start with a prior probability distribution of values of θ . This distribution may represent sheerly our degrees of belief about the various values, or it may in some way represent the results of previous research. When multiplied by the probability of the observations on the specified hypothesis and normalized to sum to 1, the result is a posterior distribution of belief about values of θ . This posterior distribution may then be used as the prior for a new experiment.

The view of statistical inference as a progressive, ongoing process contrasts with the single-shot conception implicit in the traditional approach, where, as Jeffreys (1939/1961) puts it, hypotheses are set up like coconuts, merely to stand until hit. Since the latter excludes prior distributions (except in rare cases where a prior distribution exists in frequency form, rather than as a distribution of belief, supported or otherwise), it has to treat each new experiment formally as if it were conducted in a vacuum, without benefit of previous knowledge in the area. As Hays (1963) says,

The student of classical statistics may begin to believe that the sequence “problem-sample-computation-inference” is all there is to statistical inference. It is easy to get the feeling that no one has ever drawn a sample before and no one ever will again. Although this is a parody of the true nature of statistical inference, textbooks, including this one, quite successfully leave this impression. In contrast, no one can look into Bayesian methods, however superficially, without carrying away the impression that statistical inference is an on-going process that is never completed. (pp. 856–857).

As Hays points out, the initial prior distribution is merely a way to “get the ball rolling.” Provided that the prior is not violent in its behavior (concentrating too much probability on a particular point, for example), then any differences between priors that might have been assigned by different investigators will before long be swamped by the effects of the “objective” data, upon which they are agreed.

It is useful, in fact, to regard prior distributions as having come from a previous, fictitious experiment with certain parameters. Convenience is served by choosing priors from a family which will match the likelihood function in such a way as to simplify the mathematics; prior and likelihood functions which fit each other in this way are known as conjugate pairs. For a normal process with known variance, the conjugate prior for the mean is also normal; in this case a diffuse prior can be represented as a normal distribution with an extremely large variance. Since the variance of a sample mean is inversely proportional to the sample size N , a large variance in the prior is equivalent to a very small sample. The posterior distribution effectively weights the present sample and the previous, fictitious one by their respective sizes. Thus it is that a diffuse prior will have little influence on the posterior

probability.¹³ This principle is known (cf., e.g., Edwards, Lindman, & Savage, 1963) as the principle of stable estimation.

A sharp prior, in contrast, will be equivalent to a large previous sample and will sway the posterior assessment more. In the limit, if all the prior probability is concentrated on one point, it will be unaffected by any data whatever, since any other model is viewed as impossible. Some critics (e.g., Kyburg, 1974) have read this state of affairs as making Bayesian inference wholly vulnerable to the most irrational prejudice, but perhaps that is as it needs to be. A formal theory which would budge us from certainty in any of our positions whether we agreed to it or not would be a curious instrument, to say the least.

The opposite problem, which has been pointed out by Brilmayer (1988), is more serious: If we mistakenly assign a probability of 0 to a hypothesis *a priori*, Bayes' Theorem will never allow us to correct it, since anything times 0 is 0; and sometimes we do (legitimately) want to say something is impossible. In fact, as Tribe (1971) has argued, justice *requires* starting with a presumption of innocence, but if this were expressed as a zero probability of guilt, no amount of evidence would change it.

The posterior distribution can be used in several ways in Bayesian inference. Jeffreys (1939/1961) is critical, as we saw in Chap. 5, of the use of the tail integral for significance tests and proposes in its stead his index K : If q is the null hypothesis, q' the alternative, θ the new observations, and H the everpresent background information, then:

$$K = \frac{P(q|\theta H)}{P(q'|\theta H)} / \frac{P(q|H)}{P(q'|H)},$$

which may be read as the ratio of the posterior odds to the prior odds in favor of the null. K may of course take on any (positive) value, unlike a probability. If the data θ do not change our assessment of q , then K is 1; if they strengthen the null, K will be greater than 1; they may also weaken the null, in which case a sufficiently small K will be a criterion for rejection. Jeffreys in an appendix sets up six grades of K according to the amount of support they provide for the hypothesis. The smaller values of K correspond roughly to conventional significance levels, a result which

¹³ One of the most interesting conjugate pairs is the beta prior for a Bernoulli process. It happens that a beta distribution, defined on the interval (0,1), is uniform for $r = 1$ and $N = 2$. This feature, which is comparable to having made two previous observations, one of which was a “success,” makes it a natural for representing prior ignorance or “vagueness.” Some Bayesian statisticians have suggested, though, that the most appropriate way of representing “true” ignorance or vagueness is by $r = N = 0$, an assignment which discounts prior information entirely. (The suggestion is reminiscent of Boole’s representing ignorance by a probability of 0/0 rather than $\frac{1}{2}$.) The beta distribution for this case is U-shaped, concentrating prior probability on the extreme values of 0 and 1. This distribution may thus be suitable for those rather common situations where one or the other of these extremal values is more likely than intermediate values. As a mild example, days on which there is a total cloud cover or no clouds at all are more common, at least in certain areas, than days with about a 50% cloud cover.

Jeffreys emphasizes. He presents tables of K for several common probability distributions, and it turns out that the value of, say, χ^2 for $K = 10^{-1}$ approximates the value for $\alpha = .01$, that for $K = 10^{-2}$ approximates the 5% point, and so on. Interesting differences arise for large N , however, for sometimes outcomes with $K > 1$, supporting the null, will give $p < .01$. This circumstance is usually due to violation of the assumption of independence of the observations, yielding an artificially low error. Use of the p integral exclusively would not point up the lack of independence, and the significant result would be treated as indicating a systematic effect. The problem does not arise with K ; if both K and p are calculated, a strong discrepancy can be taken as evidence of spuriousness in the significance level.

For the purpose of emulating conventional hypothesis testing, directional tests can be achieved easily enough by using the appropriate tail of the posterior distribution as the rejection criterion. Two-tailed tests pose more of a problem; indeed it has sometimes been said that Bayesian counterparts were lacking. The problem is that under any continuous prior the null point will have only an infinitesimal probability associated with it, and hence it will also in the posterior distribution. Any point an infinitesimal distance from the null represents an alternative hypothesis. In practice the problem may be circumvented by a device which has a real basis: namely, specification of a band around the point as the null hypothesis, with a width, perhaps, of a standard error of measurement. If the posterior distribution then fails to concentrate enough probability in that interval, the hypothesis may be rejected. Hays (1963) in fact raises the question of how realistic an exact null is in the first place; or, if an exact null is appropriate, whether in that case an exact alternative may not be also.

Although Bayesian methods can nearly duplicate classical tests of hypotheses, their more natural emphasis is on estimation. Hypothesis testing was designed, after all, precisely as a means of getting around inverse probability statements; if we drop that restriction, the testing of hypotheses loses much of its *raison d'être*. Point estimation in Bayesian statistics naturally uses the most likely value of the parameter, which is the mode of the posterior distribution. Under conditions of diffuse prior knowledge, the most likely value is practically the same as the maximum likelihood estimate; Jeffreys (1939/1961) shows that the difference is of the order $1/N$. Correspondences such as this can be viewed by either side as lending some legitimacy to the other.

Similarities also exist with respect to interval estimation. In Bayesian statistics an interval estimate may be obtained simply by taking an interval around the most likely value in the posterior distribution. Such intervals carry, of course, precisely the meaning which is regularly, and erroneously, attributed to confidence intervals; for, since the posterior distribution is a distribution of belief about values of the parameter, it can legitimately be said, for instance, that there is a 95% probability that a population mean lies in a given interval. The parameter being estimated, rather than the limits of the interval, is regarded as the random variable and the object of the probability statement. These Bayesian intervals are called *credibility*

intervals, to distinguish them from the classical confidence intervals.¹⁴ Credibility intervals are neither logically nor mathematically necessarily identical to confidence intervals. Some authors (e.g., Edwards et al., 1963; Novick & Jackson, 1974) emphasize the difference; others (e.g., Phillips, 1973) emphasize the similarity between them. In general, the limits for a credibility interval may be selected in such a way as to coincide, or nearly to coincide, when the prior distribution is flat; but choosing them in that way amounts mostly to a demonstration of the flexibility of Bayesian methods, for Bayesian theory has its own criteria for estimation. In particular, it is customary for a credibility interval to be selected so as to be as short as possible, in other words to encompass the regions of highest density in the posterior distribution. For symmetrical distributions, the central interval will also be the shortest, but in nonsymmetrical distributions, such as the χ^2 used in estimating variances, the confidence interval found by cutting off the upper and lower (say) 2.5% is not the shortest. The difference here is more one of convention within the two traditions and is not dictated by the logic of the theories.

Further comparisons and contrasts with classical, Fisher-Neyman-Pearson statistics can be drawn. It was remarked earlier that since Bayesian methods include more factors in the process of inference and decision, classical procedures can generally be seen as special cases. In Bayesian hypothesis testing, for example, the criterion for rejection is a low probability given the hypothesis by the posterior distribution. The posterior distribution itself is proportional to the product of the likelihood and the prior distribution. Hence a low posterior probability can result *either* from a low likelihood (extreme significance level) or a low prior probability. So, from a Bayesian perspective, use of a fixed α implies either that the data were very unlikely under the null or that the null itself was given very little credence in the first place; but the classical methods, by excluding the latter, give us no way of distinguishing these two cases.

The classical approach to inference, of course, is structured in terms of decision, but the Bayesian approach to decision-making is set apart by its regarding the losses under the various types of errors as relevant. One common criterion is to decide on the basis of the minimum expected loss. If we let R be the loss ratio—the ratio of the loss under a Type I error to the loss under a Type II error—and O be the posterior odds ratio in favor of the hypothesis, then the minimum expected loss rule amounts to rejecting if $O \times R < 1$. Again, from a Bayesian standpoint, what classical (pre-Wald) methods do, by ignoring loss calculations and insisting on a posterior probability of less than .05 for the null,¹⁵ is equivalent to insuring a Type I error for a loss value 19 times that of a Type II error. The convention may or may not be reasonable in a given decision problem; the distinctive point of the Bayesian theory here is only

¹⁴ Many authors, including Phillips (1973) and Hays (1973), call them credible intervals instead of credibility intervals; but Good (1976) says, “For chrisake don’t call them ‘Bayesian confidence intervals,’ which is a contradiction in terms” (p. 161).

¹⁵ In the absence of a prior distribution, of course, the posterior distribution coincides with the likelihood function.

that by admitting loss ratios into the equation, it allows them to be set variably rather than treated as fixed.

For the Bayesian, the usual treatment of alternatives under the classical paradigm, particularly the Fisherian, makes the inference model suboptimal. In its most widely practiced form, the classical model consists of specifying a null hypothesis and calculating the likelihood of various outcomes under that hypothesis; low-probability outcomes are taken—extratheoretically—as tending to disconfirm. The inference is really incomplete, however, without consideration of the likelihood of the results under the alternatives. If the results obtained are just as likely—or, worse, *more* likely—under some alternative, the validity of the inference is undercut. Edwards et al. (1963) offer this illustrative example, which is a variation of what has become known as Lindley’s paradox (though, as Shafer, 1982, points out, it was Jeffreys who originally called attention to it): In a z test for a zero mean, the sample z will fall between 1.96 and 2.58 2% of the time, on the average (these are the 2.5% and 0.5% points, respectively). If the null hypothesis is false, however, instead of having some specific alternative in mind, we may be indifferent across a whole range from $z = -20$ to $z = +20$. In this case, the probability of a z in the interval (1.96, 2.58) is $(2.58 - 1.96)/40 = 1.55\%$. Hence such a z , which would lead to rejection at the 5% level in a two-tailed test, would actually favor the null more than the alternative.

In general, whereas Fisherian methods can only “weaken” the null (Bayesian term, of course), Bayesian statistics can also strengthen it. Often a set of data which would lead to classical rejection would be taken by a Bayesian as supporting the null. The difference can occur partly because classical procedures, to a Bayesian, treat the null as incredible from the start anyway. Lindley (1957) demonstrated the ultimate contrast by showing how a set of data can always be constructed so as to lead to rejection of the null, at however extreme a significance level is desired, but for which the posterior odds in favor of the null, for a Bayesian, can be made arbitrarily high. As in Jeffreys’ approach to significance testing, some of the prior probability is concentrated on a point value; then the paradox arises because the Bayesian analysis is based on the posterior probability of the hypothesis—on the *ordinate* of the curve—whereas the classical criterion is the *integral* over the tail of the curve, and these two quantities behave differently at large N . From a frequentist point of view, on the other hand, the problem is just that focusing attention on the boundary points of the rejection region reveals a high proportion of incorrect decisions there (Rosenkrantz, 1973).

Of course, the prime advantage of the Bayesian approach, in the eyes of many of its practitioners—and an enviable advantage for not a few of its critics—is that it allows us to say things about what we are studying. These are almost Novick’s words:

The major problem with both Neyman-Pearson and Fisherian classical statistics is that neither system makes it possible to construct direct probability statements about the parameter of interest. Using classical statistics, we can talk about the probability of a random interval covering the mean, but we cannot talk about the probability of the mean being within a specified interval. This seems very peculiar indeed. *We are not to be permitted to make statements about those things that are the subject of our investigation.* This is an

unsatisfactory state of affairs. What we want is a system that permits us to say: “The probability that the unknown proportion is greater than φ_1 and less than φ_2 is $1 - \alpha$.” But to make probability statements about φ , we must have a probability distribution for φ . In Bayesian statistics such a distribution is possible and meaningful and indeed is the central objective of a statistical investigation. (in Blommers & Forsyth, 1977, pp. 377–378; emphasis added)

8.4 Criticism

The principal criticism of Bayesian statistics is naturally their subjectivity; there are also significant issues with the measurement or allocation of prior probabilities. Before taking up these, I want to mention a very interesting criticism by Michael Oakes (1986). The question appears to have arisen, for him, from the dilemma of the Bayesian who finds, after calculating a posterior distribution of belief, that it really isn't believable. Oakes wonders why it is necessary, after all, for a Bayesian to go through the process of probability estimation before seeing any data, then combining this distribution with the likelihood of the data to achieve the posterior distribution of belief: Why not just wait till the data are observed and construct a subjective posterior distribution directly? The question is not without practical relevance. Tribe (1971) points out that use of Bayes' Theorem in a criminal trial would force us to begin with an assessment of the probability that the accused was guilty. But concentrating on that assessment conflicts with the mental set we are legally obliged to maintain, which is the presumption of innocence. Deferring all judgment until the evidence was presented would be the only way of rendering justice. The strongest objection to Oakes' hypothetical proposal appears to be merely that people are poor estimators of probability (see Chap. 10), but that objection applies just as well, of course, to the prior distribution. It looks to me as if Oakes' simple question goes to the heart of Bayesian procedures, and I am not sure a good answer can be given.

8.4.1 *The Logical Allocation of Prior Probabilities*

The allocation of prior probabilities in Bayesian inference could be taken as an interesting illustration of the relevance of theory to practice. As a theoretical justification of Bayesian probability, the behaviorist, betting-ratio approach has held predictably more appeal to writers of textbooks for American psychologists (e.g., Novick & Jackson, 1974; Phillips, 1973) than has the rationalist theory of Jeffreys. At the point of application, however, the device of betting ratios is palpably silly and dispensable, and practicing Bayesians operate in the manner of Jeffreys (1939/1961) or Jaynes (1976) instead. There are several specific issues of concern in Jeffreys' theory, though, which thereby become relevant for most Bayesian work.

The starting points for Jeffreys are the principle of indifference and his simplicity postulate. The former is introduced as follows:

At any stage of knowledge it is legitimate to ask about a given hypothesis that is accepted, “How do you know?” The answer will usually rest on some observational data. If we ask further, “What did you think of the hypothesis before you had these data?” we may be told of some less convincing data; but if we go far enough back we shall always reach a stage where the answer must be: “I thought the matter worth considering, but had no opinion about whether it was true.” What was the probability at this stage? We have the answer already [i.e., from his Axiom 1]. If there is no reason to believe one hypothesis rather than another, the probabilities are equal. . . . To take the prior probabilities different in the absence of observational reason for doing so would be an expression of sheer prejudice. The rule that we should then take them equal is not a statement of any belief about the actual composition of the world, nor is it an inference from previous experience; it is merely the formal way of expressing ignorance. (pp. 33–34)

If the prior probability is distributed equally over possible alternatives, however, we shall never get off the ground, as Broad (1918, 1920) pointed out (Chap. 4). Jeffreys’ answer to Broad’s criticism of Laplace is contained in his simplicity postulate, a sort of stand-in for the principle of the uniformity of nature (which he rejects).

An adequate theory of scientific investigation must leave it open for any hypothesis whatever that can be clearly stated to be accepted on a moderate amount of evidence. It must not rule out a clearly stated hypothesis, such as that a class is homogeneous, until there is definite evidence against it. Similarly, it must not rule out a quantitative law stated in terms of a finite number of parameters. But this amounts to enunciating the principle: *Any clearly stated law has a positive prior probability, and therefore an appreciable posterior probability until there is definite evidence against it.* This is the fundamental statement of the simplicity postulate. (p. 129).

By the simplicity postulate, not all possible values *are* equally likely, contra Laplace. We favor “simple” values, in particular.

Jeffreys does not claim that the simplicity postulate can ever be given completely formal expression, but he does contend that it will suffice as a practical guide in actual problems. He follows Fisher in calling the assertion of a zero value the null hypothesis; the claim that a new, nonzero parameter is needed to characterize the data he calls the alternative hypothesis. His distinctive contribution at this point is, by the principle of indifference, to say that if we are undecided whether an adjustable new parameter is needed, we should concentrate half the prior probability on the null hypothesis and spread the rest evenly over the range entertained for the variable alternative. Hence simple hypotheses are favored until a substantial weight of evidence accumulates against them. If the null should be rejected, then the data can be used to estimate the value of the new parameter.

The simplicity postulate has sometimes been attacked on the ground that many hypotheses of modern science—for example, Schrödinger’s equation—are so complex they could never have acquired any appreciable probability on Jeffreys’ program. Jeffreys’ answer to this criticism is that the field of alternatives at any point is restricted to those which can (more or less adequately) accommodate all the prior data; Schrödinger’s equation may well indeed have been the simplest of the survivors. It is, in general, difficult to construct alternatives to a scientific theory which are consistent with all existing data.

Jeffreys' answer here is relevant to another issue, concerning the possible existence of unforeseen alternatives. It has sometimes been supposed that the impossibility of being sure all possible alternatives have been considered makes any assignment of prior probabilities tenuous and presumptuous at best. But unforeseen alternatives become relevant only when they lead to consequences different from those predicted by the accepted theory. If the "true" hypothesis is more complex than the current one, there is no harm in the fact that it will not be discovered until a situation arises where it leads to divergent predictions.

Of more widespread concern than the simplicity postulate itself is Jeffreys' particular manner of allocating prior probabilities. There are several issues here.

The first is mathematical; it concerns the uniform distribution over an infinite or semi-infinite range. If we distribute prior probability uniformly over an infinite range, then by Jeffreys' Convention 3 the density at any point can only be infinitesimal; and the probability that the parameter, say θ , lies within any finite limits, say in the interval (a, b) , will be zero, because that portion of the continuum will be swamped by the infinite extent beyond the interval. Similarly, if we are estimating a positive quantity, like a variance, and distribute the prior probability over the range from zero to infinity, then the probability that θ is less than any finite quantity a will be zero. This result seems to imply that the variance is necessarily infinite, in contradiction to our initial statement that we knew nothing whatever about its value.

One solution which Jeffreys considers is to abandon Convention 3 and allow the probability over an infinite range to be infinite—in other words to use infinity instead of 1 to denote certainty. Convention 3 is after all a convention, and there is no harm in adopting a new value if we stick to it consistently. The real problem with taking infinity to denote certainty, however, as Jeffreys himself points out, is that *all* nonzero probabilities then become infinite and their ratios become meaningless.

Jeffreys criticizes his own arguments so conscientiously in the course of developing them that it is difficult to keep track of where he ends up. Hacking (1965) reads him as ultimately abandoning Convention 3 in these cases. Then his criticism is well placed: that Jeffreys goes ahead to plug these distributions, integrating to infinity, into formulas which yield probabilities integrating to 1; the scale factors are omitted, and Jeffreys does not acknowledge the switch. But the operation, whatever the answer it gave, would be meaningless.

Jeffreys can also be read, however, as ultimately opting for retention of Convention 3. His arguments in support of it for the infinite case are a little slippery; they seem to amount principally to the claim that in practical cases we assume finitude and then see what happens as we allow the limits to tend to infinity. Here, as in other areas, Jeffreys, having carefully considered arguments pro and con, ultimately glosses the issue by appealing to what the practical scientist would do, and this maneuver is bound to leave philosophers of science frustrated.

It may well be safe to regard the issue as academic, however, since in practice it is hard to imagine situations arising where we should allow values over an infinite range. E. T. Jaynes (1976), a follower of Jeffreys', gives some examples. If the quantity in question is a measured length of some material object on the earth, we know that the prior density should be zero outside the range from 10^{-8} cm to 10^9 cm,

from the size of an atom to the size of the earth. The measured breaking stress of some structural material must lie between 1 dyne/cm², the pressure of sound waves due to conversation, and 10¹⁴ dynes/cm², which is 1000 times the tensile strength of any known material. And so on.

We may consider now two problems with Jeffreys' distribution of prior probability which may have more to do than the preceding with the unpopularity of his theory. The first is obvious enough to require little discussion: It is simply that the arbitrariness of Jeffreys' principles of prior probability assignment is utterly explicit. Prior probabilities can always be assigned in such a way as to produce whatever result might be desired, and indeed it sometimes seems as if Jeffreys is choosing his distributions so as to bring his results into conformity with those of Fisher, Neyman, and Pearson. That they are plausible does not alter the fact that they are arbitrary, or subjective, in the sense of being up to us.

In spite of their being subjective or arbitrary in the sense just indicated, Jeffreys takes probabilities to be *logical*: Since they are "up to us," we are obliged to be responsible investigators in *determining* the proper assignment; prior probabilities, like any others, are not merely to be stipulated, but must represent an adequate expression of our actual state of knowledge or ignorance. Once again, Jeffreys allows himself a hedge: If it is difficult or impossible (as it typically is) to know the *correct* prior distribution *precisely*, it makes little difference anyway, for the effect of any "smooth" change in the prior distribution is roughly comparable to the difference it makes to add one extra observation. But the logical view of probability, with Jeffreys as with Keynes, entails that there is in fact one correct value for a probability; Jeffreys' refusal to make a fuss over the exact value does not seem fully to take us off the hook; and probability is left as a somewhat inscrutable quantity.

Twentieth-century appraisers have been especially apt to assume that logical prescriptions for prior distributions could not be filled. Jaynes (1976) has argued that prior distributions can be derived on logical grounds, and he has done such work himself, using Shannon's information theory to get priors for problems in statistical mechanics. But if his work holds the answer to the charge that logicality is a useless criterion, it is not yet generally known. Shafer (1982) has investigated the feasibility of using empirical prior distributions, but these are usually unavailable without additional expensive research, and his own efforts to construct one suggest that such distributions may often be lumpy rather than flat, as we tend to assume. Difficulties with both logical and empirical priors have thus left most Bayesians favoring a behavioral, rather than a logical, interpretation of degrees of belief.

8.4.2 *The Subjective Allocation of Prior Probabilities*

A fundamental feature of the personalist Bayesian approach is the model of the rational decision maker—someone who faces a discrete, fully specified set of alternatives in the world, with a repertory of specifiable behaviors, the consequences of each of which, under the various possible conditions, can be expressed in profit and

loss terms. A rule for calculation of expectation must be specified (e.g., quadratic loss with distance of outcome from expectation, in multidimensional space). The individual must be prepared to place bets on all possible contingencies, and in such a way as to be protected from certain loss. The rational principle of action is the maximization of expected utility, under all circumstances.

It is thus an important objection that such extreme rationalism is unrealistic and ultimately rather silly. Keynes' (1921/1973) example about the umbrella is perhaps to the point; so is Neyman's (1952) widely quoted example about flight insurance: If we accept (not to say “believe”) the hypothesis that the plane we are about to take will crash, we should logically cancel the trip; if we reject the hypothesis, we should not in logic buy flight insurance; yet in practice we often do both. The actions are inconsistent, yet leave us with no overwhelming experience of unreasonableness. Such arguments are receiving increasing consideration in recent years, and the allure of the rationalist model has begun to fade somewhat (cf. Harper & Hooker, 1976, *passim*).

Michael Oakes (1986) raises some interesting questions from his own experience with subjective probability assessment. He found the process of generating subjective probability distributions—e.g., “Between what limits are you 99% sure that the distance from London to Berlin lies?”—“a considerable cognitive strain,” which should not surprise us, and that awkward self-reflexive puzzles arose from attempts to correct for known tendencies to overconfidence: “If I am 95% sure that a parameter lies within these limits but I know that I have a tendency to over-confidence, how confident am I *really*?” (p. 141). Oakes is talking about the simplest cases of direct probability estimation; we should hardly expect the process of subjective probability allocation to be easier if we had to consider in addition the bets we would be prepared to accept on the various propositions.

Two questions are often raised about the appropriateness of betting in scientific contexts, the first more or less esthetic, the second logical. Many scientists are personally averse to the practice of betting (in a way that may have nothing to do with religious fundamentalism) and would hate to see it made an integral part of every investigation. It would seem, moreover, to have very little to do with the essential work of science. On a more serious level, there is something decidedly odd about structuring coherent bets on natural phenomena. The point is the same one made against the minimax criterion: We are betting against only nature; the prohibition of Dutch book, needed to insure coherence of personal probabilities, seems, as Kyburg (1974) points out, to imply an assumption of nature as diabolical, seeking to win against us.

In fact, the concept of betting figures most prominently in the formalization of the personalist theory of probability; it is a convenient way of defining personal probability, in that it happens to satisfy a standard set of axioms. In textbook presentations of Bayesian statistics, and in the actual application of Bayesian methods, however, little or no mention may be made of betting. It is possible, at any rate, to select a function to represent a distribution of belief without conceiving of the process as guardedly betting against a diabolical reality. Hence the best

defense—perhaps the only one—may be to argue that betting is essentially peripheral in scientific applications of Bayesian statistics.

The logical issue is a little more serious; it has been put succinctly by Gière (1976):

The state of a theory being true is not the kind of thing one could make the object of a bet. It would not be possible even to settle such a bet, and a bet without a possible payoff is not bet at all. Similarly, the action of accepting a theory as true is not something to which we know how to assign utilities in various possible states. (p. 65)

To suppose that we can know whether our statistical hypotheses are true or false implies a degree of realism which is not easy to square with the Bayesian instrumentalism.

A further problem with Bayesian statistics is the ambiguity of the status of personal priors. If there is indeed a distinction to be made between vague knowledge and no knowledge, Bayesian statistics would appear to blur it. A statement by de Finetti (1972) suggests that he sees them as continuous with each other: “Measurements are sharper judgments; and judgments are broader measurements; there exists nothing better to do than to utilize the available data with due regard to their precision” (p. 165). Fine (1973) thinks it inappropriate that most subjective theories of probability “do not distinguish between those prior distributions arrived at after much experience and those based on only the vaguest forms of uninformed introspection” (p. 231). They make no distinction between a probability of $\frac{1}{2}$ for getting a head with a fair coin and with a coin known to be either two-headed or two-tailed; the prior probability distributions would likewise be similar for a deck of cards we have just shuffled and a deck handed to us by a stranger. Whatever the state of our knowledge or the degree of conviction, warranted or unwarranted, it is always represented in Bayesian statistics in the same way: by a (sometimes unsettlingly) exact, detailed mathematical probability distribution. To this state of affairs, Fine (1973) objects that “It is neither rational nor wise to force what few crumbs of information we may possess about a parameter into the misleadingly detailed form of a distribution” (p. 231).

The Bayesians could claim in their defense that they are attempting to handle what the classical theory excludes, by treating irregular cases (two-headed or two-tailed coins, etc.) as equivalent to cases assessed by the principle of indifference or by relative frequencies. If this equation is questionable, it is not clear how else such cases could be handled satisfactorily in a Bayesian theory of probability. Savage (1954) considered distinguishing probability assessments according to our degree of confidence in them, but ultimately rejected the idea of attaching confidence markers to confidence markers as threatening infinite regress. He was close here to an idea that might have helped, which is the concept of weight of evidence; but, although that concept has been studied by some recent theorists, it is not yet part of mainstream Bayesian procedures.

In the midst of these criticisms, however, it is important to notice a special advantage which is won by the subjectivist perspective on probability. Both techniques relied on by objective theories for the assessment of probability—the principle of indifference and empirical relative frequencies—are beset by serious problems

which have never been adequately resolved. Frequency calculations entail the selection of a reference class, neither uninformatively wide nor uninformatively narrow, and Keynes' experience with the principle of indifference seems to have discouraged anyone else from even attempting a formulation in objective terms. It is a not inconsiderable asset to the Bayesian approach that these problems vanish as a consequence of its constructivism. The Bayesian grants at the outset that judgments of similarity, symmetry, and the like are judgments—no more, no less; that they are relative to our particular cognitive purpose; and that different groupings may be selected for different purposes.

If, for example, we wanted to show that the evaluation where all probabilities were judged equal is the only “correct” one, and that if an individual does not share it he is “mistaken,” we would first have to explain what we meant by saying that an individual evaluating a probability is judging “correctly” or is “mistaken,” then to show that the considerations of symmetry referred to necessarily imply that we must accept the hypothesis of equal probabilities if we do not wish to be “mistaken.” But any event can either happen or not happen, and in neither case can we say with what degree of doubt it was “reasonable” or “correct” to expect it in advance of knowing whether or not it occurred. (de Finetti, 1937, pp. 17–18)

Similar considerations apply to the assessment of probabilities on a relative frequency basis. De Finetti regards the choice of a reference class as arbitrary in itself, to be justified in a particular case on more or less pragmatic grounds. This inevitable element of arbitrariness is his reason for avoiding expressions such as “trials of the same kind,” or “events which can be repeated”—they imply a rigid, essential, objectively given classification, which he rejects—and for substituting his concept of “exchangeable” events.

It is also worth noting at this point that both subjective and frequency definitions of probability are parasitic on the old Laplacean definition based on equiprobable cases. The frequentists, in spite of taking the long-term relative frequency as the definition, still invariably use the classical probability (e.g., 1/6 for getting a 6 with a die) as that ideal frequency which the empirical series is assumed to approximate. The Bayesians, for their part, would appear to start from a wide-open field of possible systems, but they always define conditions of coherence in such a way that rational persons are required to accept the classical probabilities as their degree of belief. Some of the force of subjectivism is lost when we are all compelled to think the same way, and little attention is given to supporting the proposition that this system of belief is the most reasonable. The apparent implication from both theories is that it may not be possible to escape the classical definition.

8.4.3 *Subjectivity*

At once the widest and deepest issue with Bayesian statistics is their subjectivity. The problem is most conspicuous for the personalists but is also present for the logical theorists. If personalist theories could be interpreted as an intermediate step toward the development of logical theories, the situation might look more

promising; but the reverse appears to be the case, as we have seen, both logically and historically. The transition from logical to personal probability is a short one, as we saw in introducing personalist theories: If logical probability is the appraisal made by a wise person, or an appraisal from the fullest context of present human knowledge, then who is to say what this appraisal shall be? It can only be up to us to say; that fact does not make us rational; and we are transported swiftly into subjective probability.

From the other end, logical probability to a personalist represents a limiting case where appraisal of a probability meets nearly universal assent; but the large N does not change the epistemological status of the assessment. Hence it does not appear that the Bayesians can adequately respond by emphasizing their logical side; their subjectivity remains an essential and ineradicable feature. It proves, however, a surprisingly difficult target to attack. One of the reasons is that the subjective theory enjoys the advantage of omnivorousness: Whatever consideration may be put forth as reasonable, the subjective theory can absorb—give it a number and find a place for it in Bayes' formula.

One criticism, however, should be dispensed with before considering a serious Bayesian defense. Bayesian methods are sometimes opposed on the ground that if probability assessments are subjective, then we may make them any way we please, to get the answer we want. Kyburg (1974), p. 116), for instance, notes that prior distributions are sometimes defended on the basis of the plausibility of the posterior distributions to which they lead. Hays (1963, p. 858) has provided the answer to criticisms of this type: that “fudging” prior distributions is simply irresponsible behavior, which no theory of probability sanctions; and, moreover, that if we are inclined to “doctor” our results, it can be done perfectly well in classical methods; in fact, it happens all the time. Acknowledgment of the subjectivity of Bayesian methods—the fact that probability assessments are “up to us” rather than objectively given—does not constitute an invitation to irresponsibility on the part of scientists.

There are two possible defenses of the alleged subjectivity of Bayesian methods, and both of them attempt to turn the tables on frequentist theories in this issue. E. T. Jaynes (1976) asks forthrightly what is so “objective,” after all, about classical or frequency probabilities, given especially their sources in scientific work.

I am unable to see why “objectivity” requires us to interpret every probability as a frequency in some random experiment; particularly when we note that in virtually every problem of real life, the direct probabilities are not determined by any real random experiment; they are calculated from a theoretical model whose choice involves “subjective” judgment. The most “objective” probabilities appearing in most problems are, therefore, frequencies only in an ad hoc, imaginary universe invented just for the purpose of allowing a frequency interpretation. The Bayesian could also, with equal ease and equal justification, conjure up an imaginary universe in which all his probabilities are frequencies; but it is idle to pretend that a mere act of the imagination can confer any greater objectivity on our methods. (p. 209)

The Bayesians, whatever the merit of their position, cannot really be charged with naïveté on the issue of objectivity. In fact, though it comes as a surprise to some readers just being introduced to subjective probability, the personalist theory has rather solid empiricist credentials. De Finetti was primarily influenced in his thinking by Hume, the Pragmatists, and especially Mach and Bridgman. The concept of degrees of belief may sound mentalistic but is fully operationalized by means of betting ratios. He was drawn to formulate a “subjective” theory, in other words, just because he saw it as behaviorist and empirical, and as more objective than the frequency theories, which try to force probability into rationalistic molds, usually involving recourse to unobservable infinite sets. The frequentists, in their attempt to be empirical, are trapped into slavish adherence to the latest empirical results. Borel’s (1939/1952) remarks are to the point.

The probability of a single case is defined subjectively by the conditions of the bet which one is disposed to accept for or against the event. . . . The probability of a single case being defined for a given individual, objective probabilities will be defined as those whose value is the same for a certain number of individuals equally well informed about the conditions of the chance event. If this event, like the throw of dice, can be repeated a large number of times under the same conditions, the theory of repeated trials informs us that the limiting value of the frequency is equal to the probability; this gives us a verification, but not a definition; if, having thrown a die a million times (which is perhaps the maximum without wearing the die to the point of modifying the conditions of the experiment), I got a six 166215 times, I would continue to believe that the probability of a six was $1/6$, i.e. .16666. . . . and not .1662. (Borel, 1939/1952, pp. 105–106)

Some frequentists, notably Mises (1928/1957), have argued that probability theorists may be permitted the same license as any other scientists in allowing for approximations to a theoretical ideal, so that we could regard a proportion of .1662 as an acceptable approximation to a true—empirical—value of .1666. . . . But other thinkers—including de Finetti (1937), Savage (1954), and even Reichenbach (1949)—have denied them that out. De Finetti’s statement is succinct:

The analogy is, in my opinion, illusory: in other sciences we have a theory which asserts and predicts with certainty and exactitude what should happen if it was fully exact; in the calculus of probability it is the theory itself which obliges us to admit all frequencies as possible. In other sciences, uncertainty thus derives from the imperfect link between theory and facts; in our case, on the other hand, it cannot have its origin in that link, but in the heart of the theory itself. No relation between probabilities and frequencies is empirical in character, because the observed frequency, whatever it may be, is always compatible with all opinions concerning the respective probabilities; these opinions consequently can be neither confirmed nor disconfirmed, inasmuch as they contain no categorical affirmation, as for example: such-and-such an event *must* be verified or *cannot* be verified. (de Finetti, 1937, pp. 23–24)

The way out of the circle, to allow the possibility of disconfirmation, is arbitrarily to regard certain very small probabilities as in fact having a value of zero. But the Bayesians are ready for those who are tempted by this approach: for it amounts to letting in just a little bit of subjectivity, and why not embrace the subjectivity fully and consistently?

Here we meet the second Bayesian counterthrust. If objectivity consists partly of making explicit what we know, then the very explicitness of the subjective aspects of Bayesian statistics might perversely be counted in their favor. They make it harder, at least, for a scientist to evade the responsibility of an active knower. The difference between Bayesian and classical methods, the Bayesians would say, is not that the former alone are subjective, but that in the classical methods the subjective aspects are more concealed. It is relevant here to recall the numerous statements of Neyman and Pearson, especially the latter, attesting to their recognition—we could almost say “insistence”—that their procedures are not exhaustive of the process of inference. Rather, they merely attempted to formalize those aspects that were readily quantifiable, without ever denying that there were many necessary aspects to the process of inference and decision which remain ineluctably personal. Their theory appealed so powerfully to social scientists of the twentieth century, however, just because leaving the subjective aspects out made it look as if what remained was a whole, objective theory.

The issue becomes concrete in the utilization of prior information or opinion. No one objects to the use of Bayes’ Theorem when prior probabilities exist in the form of frequencies, but there are real questions about the forced quantification and inclusion of vague impressions, memories, or theoretical hunches. Classical statisticians handle the issue simply by stipulating that the sample data constitute the whole of the relevant knowledge, for even informal knowledge can vitiate classical statistical results, as Kendall’s (1949) or Oakes’ (1986) criticism of confidence intervals shows. Bayesians, however, contend that it is the rule rather than the exception that we have some kind of prior information and that objectivity is not served by techniques which require the omission of such information. Actually, they go further: Without the use of initial opinion, the choice even of a criterion is left arbitrary, beyond the theory. Neyman and Pearson (1928) considered a handful of different criteria for testing and opted finally for the p integral of the likelihood contour on the more or less arbitrary ground that it afforded the closest fit to conventional practice up to that time; Wald (1950) explicitly considered losses in decision functions but arbitrarily selected the minimax strategy, which is appropriate primarily for zero-sum games.

From a frequentist point of view, the Bayesian approach, anchoring the likelihood function in subjective prior and posterior probabilities, could be compared with fastening a steel tie-beam into plaster at both ends. The analogy is Jeffreys’ (1939/1961), however, in criticism of the frequentist approach. In a widely quoted remark, de Finetti (e.g., 1972, p. 160) says of the requirement that prior information be excluded that “This is like saying that, in view of the danger of building on sand, one need but eliminate the sand and build on the void to remove every danger.” The Bayesians claim it as their advantage that they attempt to deal with such criteria rationally and theoretically rather than leaving them as arbitrary, when certain standard choices may have dubious relevance for some applications.

Indeed, subjectivists extend the domain of objectivity by demonstrating that even the choice of a criterion is rendered unique for objective reasons, except for the arbitrariness in the initial opinion which is not objective but is, at least subjectively, significant, while those who call themselves objectivists abandon this whole field to the domain of the purely arbitrary. Moreover, as has already been said, subjectivists are required by their theory to completely utilize all objectively known facts (from whatever sort of observation or experiment), without assuming any right to manipulate or to set aside some of the data arbitrarily. I find it most strange that anyone is satisfied simply to ignore all that is subjectively meaningful and much that is objective too in order to console himself that he is progressing toward more objective thought. (de Finetti, 1972, p. 186)

The objectivists may make at least one final response to this criticism: that just as the Bayesians charge them with achieving an appearance of objectivity by hiding rather than eliminating subjective aspects of inference and decision, so they themselves may be charged with achieving an appearance of objectivity by forcing subjective hunches and the like into misleading numerical form on a par with actual frequencies; that, possible contrary impressions notwithstanding, exclusion of subjective aspects from the *theory* does not mean that they should be or could be excluded from the actual *process* of inference and decision; and that science is best served by formalizing only what lends itself well to formalizing, but maintaining an attitude of cognitive responsibility throughout all aspects of scientific work.

8.5 Putting Theories to Work

Star and Gerson (1987), among others, argue that, to understand science, we have to look at the actual *work* done by scientists. That perspective is an interesting one to apply to the theorists we have reviewed in the last three chapters.

In the first place, among the probability theorists, all of the frequentists—Venn, Mises, Reichenbach, Popper—were philosophers. (Mises also did some work in physics.) The obvious cynical rejoinder is that philosophy isn't work, and there are accordingly no real-world constraints on their theorizing. It is tempting to add that the lack shows. Virtually every scientist in the first half of the twentieth century claimed to be a frequentist—but none was an adherent of any *specific* frequency theory. The frequentist label served much as the Darwinist label functioned in biology. That label was used to identify the speaker as an opponent of creationism, which was perceived to be the only alternative. Similarly frequentism denoted merely opposition to inverse probability, rather than any specific doctrine.

Among frequentist statisticians, Neyman functioned more as a pure mathematician, whereas Fisher was active in research, in genetics and agriculture—and he used the methods he developed in his work. What is intriguing in this context about Fisher is that he was coming from the background of the biometricalians, using statistics to study biological populations—but in his agricultural work there were no populations. Fisher solved the problem simply by pretending that there were—his hypothetical infinite population—and that his conclusions pertained to that nonexistent entity.

The Bayesians are an interesting contrast. Although Keynes wasn't anything but a philosopher at the time of his *Treatise on Probability*, Jeffreys and Jaynes were both physicists, actively concerned with solving real-world problems. Howie (2002) very usefully notes the different kinds of research done by Jeffreys and by Fisher. Fisher's thinking, if not his actual work, was concerned with populations and random sampling. The uncertainties confronted by Jeffreys, as a geophysicist, didn't have to do with randomness or sampling distributions, but with sparse measurements of uncertain accuracy, of different kinds of quantities—pooling seismograph records with data on tides to make inferences about the composition of the earth's core, for example. What Howie surprisingly does not mention, however, is the very interesting fact that Jeffreys' work in geophysics gives no indication of using his own tools of Bayesian inference. There is almost no mention of statistics in his standard textbook *The Earth* (1924/1962); but neither is there in his research papers, where it would be expected (e.g., Jeffreys, 1940, 1949). It is as though all his theorizing about probability and scientific inference were an academic exercise only, sheerly for the esthetic value of formalization, with no expectation that it would ever be put into practice.

So Fisher was the only one of the lot who used his own tools—but in a way that manifestly did not match the nature of the work. One almost wonders if there is a reason for the startling gap between theory and practice.

References

- Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'école américaine [The behavior of the rational man confronting risk: Critique of the postulates and axioms of the American school]. *Econometrica*, 21, 503–546.
- Ayer, A. J. (1962). The conception of probability as a logical relation. In S. Korner (Ed.), *Observation and interpretation in the philosophy of physics* (pp. 12–17). New York, NY: Dover. (Original work published 1957).
- Bateman, B. W. (1987). Keynes's changing conception of probability. *Economics and Philosophy*, 3, 97–120.
- Bernoulli, J. (1713). *Ars conjectandi*. [The art of conjecturing]. Basel, Switzerland: Thurneysen.
- Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston, MA: Houghton Mifflin.
- Borel, É. (1924). À propos d'un traité de probabilités [Review of *A treatise on probability*]. *Revue philosophique*, 98, 321–336.
- Borel, É. (1952). *Traité du calcul des probabilités et de ses applications* [Treatise on the calculus of probabilities and its applications]. Tome 4. *Applications diverses et conclusions* [Vol. 4. Various applications and conclusions]. Fascicule 3. *Valeur pratique et philosophie des probabilités* [Book 3. Practical value and philosophy of probabilities] (2nd ed.). Paris: Gauthier-Villars. (1st ed., 1939).
- Brilmayer, L. (1988). Second-order evidence and Bayesian logic. In P. Tillers & E. D. Green (Eds.), *Probability and inference in the law of evidence* (pp. 147–167). Dordrecht, The Netherlands: Reidel.
- Broad, C. D. (1918). On the relation between induction and probability. Part I. *Mind*, 27, 389–404.
- Broad, C. D. (1920). On the relation between induction and probability. Part II. *Mind*, 29, 11–45.

- Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). Chicago, IL: University of Chicago Press. (1st ed., 1950).
- Cohen, L. J. (1977). *The probable and the provable*. Oxford, UK: Clarendon Press.
- Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. [Exposition of the theory of chances and probabilities]. Paris, France: Hachette.
- Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins University Press.
- De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives [Foresight: Its logical laws, its subjective sources]. *Annales de l'Institute Henri Poincaré*, 7, 1–68.
- De Finetti, B. (1972). *Probability, induction and statistics*. New York, NY: Wiley.
- De Finetti, B. (1974). *Theory of probability* (Vol. 1). New York, NY: Wiley.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Ellis, B. (1973). The logic of subjective probability. *British Journal for the Philosophy of Science*, 24, 125–152.
- Fine, T. L. (1973). *Theories of probability*. New York, NY: Academic Press.
- Gièvre, R. N. (1976). Empirical probability, objective statistical methods, and scientific inquiry. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 1. *Foundations and philosophy of statistical inference*, pp. 63–101). Dordrecht, The Netherlands: Reidel.
- Gillies, D. (1988). Keynes as a methodologist. *British Journal for the Philosophy of Science*, 39, 117–129.
- Good, I. J. (1950). *Probability and the weighing of evidence*. London, UK: Griffin.
- Good, I. J. (1976). The Bayesian influence, or how to sweep subjectivism under the carpet. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2. *Foundations and philosophy of statistical inference*, pp. 125–174). Dordrecht, The Netherlands: Reidel.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hansel, C. E. M. (1966). *ESP: A scientific evaluation*. New York, NY: Scribners.
- Harper, W. L., & Hooker, C. A. (Eds.). (1976). *Foundations of probability theory, statistical inference, and statistical theories of science*. Dordrecht, The Netherlands: Reidel.
- Harris, E. E. (1970). *Hypothesis and perception*. New York, NY: Humanities Press.
- Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart, Winston.
- Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, Winston.
- Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart, Winston.
- Howie, D. (2002). *Interpreting probability: Controversies and developments in the early twentieth century*. New York, NY: Cambridge University Press.
- Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2. *Foundations and philosophy of statistical inference*, pp. 175–258). Dordrecht, The Netherlands: Reidel.
- Jaynes, E. T. (2003). In G. L. Bretthorst (Ed.), *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.
- Jeffreys, H. (1940). The variation of latitude. *Monthly Notices of the Royal Astronomical Society*, 100, 139–155.
- Jeffreys, H. (1949). Dynamic effects of a liquid core. *Monthly Notices of the Royal Astronomical Society*, 109, 670–687.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press. (1st ed., 1939).
- Jeffreys, H. (1962). *The earth: Its origin, history, and physical constitution* (4th ed.). Cambridge, UK: Cambridge University Press. (1st ed., 1924).
- Kempthorne, O. (1976). Statistics and the philosophers. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2. *Foundations and philosophy of statistical inference*, pp. 273–314). Dordrecht, The Netherlands: Reidel.

- Kendall, M. G. (1949). Reconciliation of theories of probability. *Biometrika*, 36, 101–116.
- Kendall, M. G. (1968). On the future of statistics—A second look. *Journal of the Royal Statistical Society, Series A*, 131, 182–204.
- Keynes, J. M. (1933). *Essays in biography*. London, UK: Macmillan.
- Keynes, J. M. (1973). *A treatise on probability*. New York, NY: St. Martins Press. (Original work published 1921).
- Kiefer, J. (1977). The foundations of statistics—Are there any? *Synthèse*, 36, 161–176.
- Kneale, W. (1949). *Probability and induction*. Oxford, UK: Clarendon Press.
- Kyburg, H. E., Jr. (1974). *The logical foundations of statistical inference*. Dordrecht, The Netherlands: Reidel.
- Kyburg, H. E., Jr., & Smokler, H. E. (Eds.). (1964). *Studies in subjective probability*. New York, NY: Wiley.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.
- Matalon, B. (1966). Épistémologie et psychologie des probabilités [Epistemology and psychology of probabilities]. In F. Bresson & M. de Montmollin (Eds.), *Psychologie et épistémologie génétiques [Psychology and genetic epistemology]* (pp. 107–115). Paris, France: Dunod.
- McGrayne, S. B. (2011). *The theory that would not die: How Bayes' rule cracked the Enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy*. New Haven, CT: Yale University Press.
- Misak, C. (2020). *Frank Ramsey: A sheer excess of powers*. Oxford, UK: Oxford University Press.
- Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. Washington, DC: U.S. Department of Agriculture Graduate School.
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, 20A, 175–240.
- Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society*, 24, 492–510.
- Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York, NY: McGraw-Hill.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.
- Patinkin, D. (1976). Keynes and econometrics: On the interaction between the macroeconomic revolution of the interwar period. *Econometrics*, 44, 1091–1123.
- Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London, UK: Nelson.
- Polya, G. (1954). *Mathematics and plausible reasoning* (Vol. 1. *Patterns of plausible inference*). Princeton, NJ: Princeton University Press.
- Popper, K. R. (1968). *The logic of scientific discovery* (2nd English ed.). New York, NY: Harper Torchbooks. (Original work published 1934).
- Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society, Series B*, 27, 169–203.
- Ramsey, F. P. (1931). *The foundations of mathematics*. London, UK: Kegan Paul, Trench, Trubner.
- Reichenbach, H. (1949). *The theory of probability* (2nd ed.). Berkeley, CA: University of California Press.
- Rosenkrantz, R. D. (1973). The significance test controversy. *Synthèse*, 26, 304–321.
- Rothbard, M. N. (1992). Keynes, the man. In M. Skousen (Ed.), *Dissent on Keynes: A critical appraisal of Keynesian economics* (pp. 171–198). New York, NY: Praeger.
- Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley.
- Schlaifer, R. (1959). *Probability and statistics for business decisions*. New York, NY: McGraw-Hill.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, 77, 325–351.
- Spencer Brown, G. (1972). *Laws of form*. New York, NY: Julian Press.
- Star, S. L., & Gerson, E. M. (1987). The management and dynamics of anomalies in scientific work. *Sociological Quarterly*, 28, 147–169.
- Todhunter, I. (1865). *A history of the mathematical theory of probability from the time of Pascal to that of Laplace*. Cambridge, NY: Macmillan.

- Toulmin, S. (1964). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, 84, 1329–1393.
- von Mises, R. (1957). *Probability, statistics and truth* (2nd English ed.). New York, NY: Macmillan. (Original work published 1928).
- Wald, A. (1950). *Statistical decision functions*. New York, NY: Wiley.
- Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 38, 715–731.

Chapter 9

Statistical Inference in Psychological and Medical Research



The preceding chapters have shown how the modern theory of statistical inference arose from the large-sample biometric research of Karl Pearson's group and from agricultural field trials in the work of R. A. Fisher. At the turn of the twentieth century, it would not have been obvious how either of these might have been extended to the psychological science of the day. Several further technical developments were required before statistical inference could take its place in psychological research in a way that would be recognized today. (a) One, obviously, was psychological measurement. (b) Another was the concept of the treatment group. (c) A more subtle conceptual shift was the transition from large-sample to small-sample theory. Application of statistical inference to psychological research gave rise to a number of practical issues and controversies, which we shall consider in this chapter, along with possible responses by Bayesians. The history of Bayesianism in psychological research will be reviewed, from its introduction in 1963 to recent inroads due to certain technical advances, notably the multiple imputation of missing data. Finally, we shall briefly consider the rather different history of the incorporation of statistical inference into medical research and some of its consequences.

9.1 Psychological Measurement

The step from biometrics to psychometrics was a short one, at least to retrospect, but it presupposed that psychological phenomena could be measured. That premise has had a long, fascinating, and continuing history, chronicled by Hornstein (1988). Before looking at the development of psychometrics, however, I want to consider briefly the parallel developments in econometrics. We saw in Chap. 2, after all, that economics, in particular the market price system, appears to have been the origin of

the idea that all qualities could be numerically measured.¹ The thirteenth- and fourteenth-century philosophers of money, however, had an insight which was discarded by their counterparts in the nineteenth and twentieth centuries.

If I buy, say, a car for \$4450 (the price of the last new car I bought), it is because, at this moment, I prefer the car to the \$4450. And the dealer sells it because he prefers the \$4450 to holding onto the car. Both parties perceive themselves to gain by the transaction. The dealer doesn't know how much more I might have paid for it, and I don't know how much less I might have been able to get it for. The fact that the "market price" of this transaction thus represents a range of possible values at which it might have taken place was recognized by medieval philosophers in the concept of *latitudo* (Kaye, 1998), for instance, in the writings of Duns Scotus and Pierre Olivi.² Aristotle had conceived of the just price as a point and had thought that it was often necessary for a neutral, third-party human judge to intervene to set the just price between competing interests of buyer and seller. With the tremendous expansion of monetization and commerce in the thirteenth century, Aquinas acknowledged that the just price couldn't be limited to a single point:

The just price of things sometimes is not precisely determined [*quandoque non est punctualiter determinatum*], but rather consists in a certain estimate [*quadam aestimatione consistit*]. Therefore a small addition or subtraction does not seem to destroy the equality of justice. (quoted in Kaye, 1998, p. 99)

Equality came about as an *equality of doubt* (Kaye, 1998). In the formulation of Buridan a century later, equality emerged as a *product of willed inequalities*. The process of voluntary exchange itself seemed to guarantee perceived equality, obviating Aristotle's third-party judge. No one in the fourteenth century yet explicitly identified the just price with the market price, for the former was an ethical concept, which belonged to the realm of universals, whereas the latter was a matter of messy empirical particulars. Money was still seen as too inherently corrupt (St. Francis reportedly refused to touch it) to constitute a model of scientific measurement. But the fourteenth century was nevertheless acquiring a new comfort with the concepts of estimation and approximation (Kaye, 1998).

Theocharis (1961) finds a number of economists, especially Italians, applying mathematics to economics, in a way that implicitly takes prices to be point values, from the eighteenth century on. An interesting exception is Nicolas François Canard, a mathematician rather than an economist, who wrote in response to a prize offered by the *Institut National des Sciences et des Arts* on the question of whether, in an

¹ Isidore of Seville argued, in the seventh century, that *number* derives from the Latin *nummus*, coin (Kaye, 1998). Partridge (1958/1966) supports at least a connection.

² Olivi, a Franciscan from Narbonne, is less well known than his thought deserves. Having formulated the concept of subjective utility, thereby providing a better solution to the "value paradox"—that water, which is essential for life, should be cheaper than diamonds, which are not—than was available for another 600 years, he was also the leader of the Spiritual branch of the Franciscans, who took seriously the vow of poverty. That sufficiently threatened the Conventional Franciscans that after his death, in 1298, they ordered his body exhumed and his bones scattered, and his works destroyed (Rothbard, 1995/2006a, v. 1).

agricultural economy, all taxes fell ultimately on land. Canard (1801) explicitly retained the medieval term *latitude* for the range between the minimum and maximum price. It seems likely that he picked up the term from Scotus or Olivi, though, as was the custom of the day, he acknowledged no sources. Canard won the prize—and the resentment of all the more established economists in France. He had some influence on the Italians, but was still being denounced by Schumpeter 150 years later. So his fame unfortunately stifled his influence.

Recognition that prices didn't constitute point measurements persisted at least as late as 1837, in the writings of William Lloyd, a minor Tory figure:

It would indeed be difficult to discover any accurate test, by which to measure either the absolute utility of a single object, or the exact ratio of the comparative utilities of different objects. Still it doesn't follow, that the notion of utility has no foundation in the nature of things. It does not follow, that because a thing is incapable of measurement, therefore it has no real existence. The existence of heat was no less undeniable before thermometers were invented, than at present. (quoted in Rothbard, 1995/2006b, vol. 2, p. 129)

But mathematical treatment, in macroeconomics and econometrics, required yielding to the seduction of prices as point values—ones which, moreover, were fixed rather than constantly varying across place and time. Credit for the first major equation of economics, the “equation of exchange,” is generally attributed to Irving Fisher (1913), though Fisher himself presents it, in textbook fashion, as an established relation, requiring no particular justification, and he gives priority to other authors, including Edgeworth in 1887.³ In the equation of exchange, the general price level P is said to be determined by three factors: the quantity of money in circulation, M ; its average velocity of circulation, V ; and the total quantity of goods sold, T : $MV = PT$. In Fisher's example, someone buys 10 pounds of sugar at 7¢ a pound. From this transaction, Fisher claims that “10 pounds have been regarded as equal to 70 cents” (p. 16). But neither the buyer nor the seller regarded them as equal: The buyer regarded the 10 pounds of sugar as more valuable than the 70 cents and vice versa for the seller. Fisher then goes on to claim that the equation of exchange for the economy as a whole is obtained by adding up all such individual transactions for a specified period of time. So let us add a second consumer who buys a book for \$10. The total money spent, which Fisher calls E , is \$10.70, which obviously equals the total money received. But this quantity cannot be factored into PT so easily as Fisher imagines. To arrive at P , we have to divide \$10.70 by 10 pounds of sugar plus a book. But sugar and books are incommensurable (How should the cost of a concert be reckoned per pound?); they cannot be added. So P , the value to be “explained” by the other factors in the equation, can't be calculated, separate from the product PT . Similarly for V on the other side: If, in a given hour, I have \$200 on hand, and spend \$10 on a book, my V is 1/20. V cannot be defined

³That Edgeworth should have pioneered a formula pertaining to values is no surprise, considering that he had written a whole book on the mathematical analysis of pleasure (cf. *infra*, Chap. 1). How Fisher gained access to Edgeworth's report I am not sure. The library of the University of California, San Francisco, informed me that the only library in the world which held this report was the House Senate Library in Britain, which does not lend or photocopy.

independently of M . So at best the vaunted equation of exchange expresses only the truism that money spent = money received. There is no economic insight or explanation in the fictitious concept of a general price level (cf. Rothbard, 1962/2009).

It was a mere quarter century after Fisher modestly put forward the equation of exchange that the Dutch economist Jan Tinbergen introduced the first econometric model, for the purpose of assessing theories of the business cycle. This work was undertaken on behalf of the League of Nations, and it earned Tinbergen the first Nobel Prize in economics (shared with Ragnar Frisch, who coined the terms *econometrics* and *macroeconomics*).⁴ Tinbergen undertook to infer, from multiple regression models, which specified predictors were more important. The most interesting aspect of Tinbergen's project, for our purposes, is the critique it received from none other than John Maynard Keynes, whose *The General Theory of Employment, Interest, and Money* (1936) is commonly considered to have launched the field of econometrics.

It was clear in Chap. 8 that Keynes was extremely conservative, and skeptical, on the issue of measurement of probabilities. Indeed he is the only major theorist to have denied that all probabilities were measurable, or that they could be ordered. That claim cost him dearly, in terms of the total neglect of his *Treatise on Probability*. To most of his readers, his stance seemed idiotically self-defeating: Without numbers, there was no way to get on with the business of playing scientist. Fifteen years later, writing *The General Theory*, he was no less skeptical of quantification in economics than in logic. But he was also eager not to have his theory neglected. The solution he seized on could have been successful only in the hands of someone as cynically supercilious as Keynes: Alongside the rather limited mathematics, he left in the scathing critique of mathematical analysis, as if daring readers to take him seriously. Macroeconomic concepts like Fisher's general price level he saw as material for historical, but not quantitative causal analysis:

The fact that two incommensurable collections of miscellaneous objects cannot in themselves provide the material for a quantitative analysis need not, of course, prevent us from making approximate historical comparisons, depending on some broad element of judgment rather than strict calculation, which may possess significance and validity within certain limits. But the proper place for such things as net real output and the general level of prices lies within the field of historical and statistical description, and their purpose should be to satisfy historical or social curiosity, a purpose for which perfect precision—such as our causal analysis requires, whether or not our knowledge of the actual values of the relevant quantities is complete or exact—is neither usual or necessary. (1936, pp. 39–40)

Speaking, in one of the few places where he indulges in equations, of the Quantity Theory of Money $D = MV$ (where D = demand), and considering the possibility of variable velocity, he says:

I do not myself attach much value to manipulations of this kind; and I would repeat the warning, which I have given above, that they involve just as much tacit assumption as to

⁴Tinbergen's younger brother Nikolaas is better known to psychologists for his work in ethology, for which he received the Nobel Prize in physiology, making the Tinbergens the only sibling pair to have received Nobel Prizes.

what variables are taken as independent (partial differentials being ignored throughout) as does ordinary discourse, whilst I doubt if they carry us any further than ordinary discourse can. (p. 305)

His disparagement of quantification generally in economics was quintessentially Keynesian:

To say that net output today is greater, but the price-level lower, than ten years ago or one year ago, is a proposition of a similar character to the statement that Queen Victoria was a better queen but not a happier woman than Queen Elizabeth—a proposition not without meaning and not without interest, but unsuitable as material for the differential calculus. Our precision will be a mock precision if we try to use such partly vague and non-quantitative concepts as the basis of a quantitative analysis. (1936, p. 40)

To pull it off, he also made the text impenetrably obscure; even his most fervent admirers, like Samuelson, admit that it is a very difficult read. It was a book which no one wanted to read, but everyone wanted to be thought to have read. So the inconsistencies, if they were noted at all, were apt to be treated as quirky manifestations of Keynes' genius. Psychology has never had a critic of quantification as esteemed as Keynes; Keynes' example shows that it wouldn't have made any difference if it had.

In reviewing Tinbergen's quantitative innovations, however, Keynes could voice his skepticism without restraint. Keynes notes at the outset that:

Prof. Tinbergen is obviously anxious not to claim too much. If only he is allowed to carry on, he is quite ready and happy at the end of it to go a long way towards admitting, with an engaging modesty, that the results probably have no value. The worst of him is that he is much more interested in getting on with the job than in spending time in deciding whether the job is worth getting on with. (Keynes, 1939, p. 559)

Among the points Keynes makes are the following:

1. Tinbergen's method is relevant only to situations where all the relevant factors can be measured. If this condition is not met, then:

it withdraws from the operation of the method all those economic problems where political, social and psychological factors, including such things as government policy, the progress of invention and the state of expectation, may be significant. In particular, it is inapplicable to the problem of the Business Cycle. (p. 561)

Perhaps inspired by Irving Fisher, Tinbergen "insists that his factors must be measurable, but about the units in which he measures them he remains singularly care-free, in spite of the fact that in the end he is going to add them all up" (p. 563).

2. The variables in the model must be exhaustive of the list of significant causes.

[T]he method is only applicable where the economist is able to provide beforehand a correct and indubitably complete analysis of the significant factors. The method is one neither of discovery nor of criticism. It is a means of giving quantitative precision to what, in qualitative terms, we know already as the result of a complete theoretical analysis. (p. 560)

Much the same point had been emphasized by Sewall Wright (1921) in his introduction of path analysis.

3. Tinbergen does not discuss whether the predictors must be independent. That will not generally be the case in economic analysis. But dependent predictors often lead to spurious correlations:

[I]f what is really the same factors is appearing in several places under various disguises, a free choice of regression coefficients can lead to strange results. It becomes like those puzzles for children where you write down your age, multiply, add this and that, subtract something else, and eventually end up with the number of the Beast in Revelation. (p. 562)

4. Tinbergen assumes throughout that correlations are linear.

But it is a very drastic and usually improbable postulate to suppose that all economic forces are of this character, producing independent changes in the phenomenon under investigation which are directly proportion to the changes in themselves; indeed, it is ridiculous. (p. 564)

A particular problem is posed by the cyclical nature of the dependent variable: It is not obvious how a cyclical pattern can be generated unless the predictors themselves are cyclical.

5. Tinbergen's attention is limited to the data at hand, in the manner of descriptive curve-fitting. But the use to which the League of Nations and others would be eager to put his results are a matter of inductive generalization, which would require the assumptions of uniformity and homogeneity. For this purpose Keynes thinks a necessary first step would be to check whether a regression coefficient obtained from a 40-year period also obtains for the component decades analyzed separately. But Tinbergen showed no interest in such questions.

The relevance of Keynes' critique here to the modeling of psychological phenomena, including structural equation modeling, surely requires no elaboration. There is no claim here that econometrics either preceded or influenced psychometrics. Roots of psychological measurement, as we shall shortly see, went back to the mid-nineteenth century, and Spearman had already developed factor analysis in 1904—a procedure to which T. W. Anderson, in his classic *An Introduction to Multivariate Statistical Analysis* (1957), devoted the enormity of a page—the proportion of unknowns to knowns was high enough to give factor analysis an indeterminacy which, like the children's puzzles Keynes (1939) alluded to, promised little of scientific value. Rather, both psychometrics and econometrics surely sprang from the giddy prospect, in the late nineteenth and early twentieth centuries, of elevating the “social sciences” to scientific status.

Stigler (1986) observes that the discovery of the personal equation in astronomy, at the end of the eighteenth century (Chap. 5), created an opening for the development of psychological measurement, but it was passed by. The question of psychological measurement remained where it had been 600 years earlier—an academic curiosity (in the worst sense of the term), with little promise of practical importance. And the practical “need” that was perceived in the United States in the twentieth century, which spurred such energetic work, was different from the perceived need which motivated the first stages in nineteenth-century Germany. The issue there was to distinguish the emerging discipline of psychology from the philosophical focus

from which it had sprung, and to endow it with the prestigious status of an empirical science. (A similar disciplinary struggle took place in economics at the same time, as the historicist school gave way to the new mathematical-statistical discipline in the *Methodenstreit*; Mises, 1966, 1969/2007.) Boring (1966) contends that Fechner's personal motivation for his work in psychophysics was his interest in refuting the materialism that was dominant at the time, even in Germany; he hoped to accomplish that by showing that psychological and physical phenomena were related by a quantitative law, thus giving validity to the former.

The line of work proceeding from Fechner contributed significantly to the eventual acceptance of the idea that psychological phenomena could be quantified. His psychophysical research was the only way in which statistics had yet entered the science of psychology, and his use of the Gaussian model of errors of measurement conceived errors in the original sense of true errors in the measurement of some single putative quantity, rather than in the later, Galtonian sense of variations among individuals. As Danziger (1987) observes, the object of psychological inquiry was at that time psychological processes, and research was typically conducted with a single individual, or a small group of colleagues. (Fechner was experimenting on himself.) Statistics as a science of aggregates would have no place in a discipline so construed.

Fechner was concerned with the measurement of sensation. Converting Weber's concept of a just noticeable difference from an observational to a theoretical unit, he assumed that all jnds were equal in subjective magnitude and thus could be treated as units. On this basis he presented his famous $\psi = k \log \varphi$, where φ is the magnitude of sensation in jnd units and ψ is a physical magnitude in appropriate physical units. Fechner (1860/1966) compared what he called Weber's Law with Daniel Bernoulli's solution to the St. Petersburg problem (Chap. 3): As Bernoulli had argued that the value of a ducat was (inversely) proportion to how many we already had, Fechner noticed that a jnd depended on the quantity assessed. We could notice a difference of half an ounce in assessing a weight of 1 ounce, but not in a weight of 50 pounds. Hence the same logarithmic law held. Actual data were to be determined by his three methods of limits, of average error, and of constant stimuli. (The names vary; Fechner, 1860/1966, called them the method of just noticeable differences, the method of right and wrong cases, and the method of average error.) Interestingly, Fechner's work seems implicitly to assume a *latitude* for the numerical value of a sensation. His method of limits mimics an auction in converging on the maximum price.

Hornstein (1988) finds that Fechner's work was controversial, the most important issue being the so-called quantity objection. William James gave it its most famous statement in his claim that the experience of pink was not some fraction of the experience of red. Although Fechner claimed to be measuring sensation, in other words, there was some question whether he was really just measuring stimuli. As Hornstein notices, this debate was never really resolved, but was merely side-stepped. The quantity objection was discredited, if not really refuted, by its appeal to introspection, which was losing ground to the rising empiricism in psychology. Fechner's followers found, meanwhile, that they could use his methods to gather

perceptual data without having to provide a theoretical justification for the idea that sensation could be measured. Hornstein argues that this split between theory and practice helped to support the idea of methods as theoretically neutral.

A similar development took place in the field of mental testing, though the method of quantification differed. The source lay here in the work of Galton and Pearson, particularly in the idea that individuals could be measured by their relative position in a distribution which was assumed to be normal. Pearson thought that mental and moral characteristics ought to be hereditary, like biological characters, and constructed his own crude 7-point scale of intelligence to investigate the hypothesis. Binet's intelligence scale retained the principle of measuring not absolute intelligence, but rather a child's position relative to others of the same age.

The field of mental measurements confronted essentially the same problem as psychophysics: It had to overcome the traditional view that intelligence was not susceptible to quantification and to demonstrate what it was that was being measured. Neither of these challenges, according to Hornstein, was met. Psychologists were too busy constructing and administering tests to pause very long to consider what, if anything, they were measuring. The new empiricist philosophy of operationism also helped: Here began the notion that intelligence is what intelligence tests measure. The logically prior question of what qualified as an intelligence test was lost; instead, attention focused on the technical question of comprehending the diversity of putative measures of intelligence through methods such as factor analysis. Thus, in much the same way as in psychophysics, Hornstein argues, sweeping substantive issues under the carpet made methodology look theoretically neutral. And the question of what, if anything, intelligence tests measure has not even seemed pressing until the cultural challenges of the last few decades.⁵

The hitherto disparate domains of psychophysics and mental measurements were united, according to Hornstein, by Thurstone's and Guilford's work on psychometric scaling. To establish scale points in an attitude⁶ variable, Thurstone used a separate sample of 300 readers. Once he had compiled about 100 statements representing some degree of the attitude to be measured, he had the readers sort them in 11 piles, from least to most representative. Ideally, readers all across the spectrum in their own views could at least agree on which of two statements was more favorable to,

⁵To my mind the most striking finding with respect to intelligence tests is that uncovered by feminist scholars in the 1980s (Jaggar & Bordo, 1989). Going through Lewis Terman's papers at Stanford, they found that Terman and his team had had to adjust the items to bring women's IQs down to the level of men's. Presumably a gender difference in the opposite direction would not have been felt to require such an adjustment—just as obtained racial differences, given their direction, were not deemed to call for any correction. Could there be any starker illustration of the arbitrariness of psychological measurement, its utter dependence on the presumptions of the test constructors?

⁶Thurstone conceived of an attitude as the dimension underlying various items as verbal expressions of opinion. He gave this striking example: “If a man says that we made a mistake in entering the war against Germany [i.e. World War I], that statement will be here spoken of as an opinion. . . . Our interpretation of the expressed opinion is that the man's *attitude* is pro-German” (1928, p. 531). In Thurstone's mind, evidently, the only basis for opposing any war was favoring the other country over the United States.

say, alcohol prohibition. Distances between scale points, say a and b , could be determined from the proportion who agreed in ranking b higher than a . Theoretically, the distance would also depend on the respective item variances and their correlations. A halo effect might cause successive items to be correlated positively; a contrast effect, where, say, a particularly good handwriting specimen made the next one look particularly bad, would induce a negative correlation. But that many parameters made the scaling problem insoluble; and Thurstone argued that it would often be feasible to treat all interitem correlations as 0, and all item variances as the same—though, he said, these assumptions should be tested.

So cumbersome a procedure of course begged for simplification, and it was only 3 years in coming. Rensis Likert (1932) sought to dispense with the readers ranking items into piles.⁷ He simply assigned the numbers 1–5 to response options from *Strongly disagree* to *Strongly agree*, then, assuming a normal distribution, converted these to z scores. But he subsequently found that, if he used the raw 1–5 scores, they correlated .99 with results using z scores. And, on the Thurstone-Droba War Scale, scoring by Likert's 1–5 method correlated .83 with the original results.⁸ With Likert's simplification of Thurstone's method, the way was open for quantification of virtually any subjective dimension.

Meanwhile, Hornstein notes ironically, the debate about measurement had been reopened in psychophysics, where it was becoming evident with the collection of more data that the various psychophysical methods yielded inconsistent results. “A committee of the British Association for the Advancement of Science considered the issue for eight years throughout the 1930's, and being unable to reach any resolution, was finally discharged” (Hornstein, 1988, p. 28). Many psychologists would consider the issue to have been resolved by S. S. Stevens (1951), with his theory of scale types and permissible statistical operations. The extent of Stevens' victory, at least in a political sense, can be judged from the fact that many of those who have rejected his conclusions have nevertheless accepted his premises. There is scarcely a statistics textbook which does not include a definition of his four scale types.

In the judgment of some of his critics, however, Stevens' formulation was unduly shaped by his commitment to the project of psychophysics. In its quest for a mapping of sensory onto physical quantities, psychophysics presses powerfully toward a representational theory. Yet Adams (1966) argues that a representational theory makes sense, if at all, only for what he calls intrinsic measurement. The Mohs hardness scale, for example, is extrinsic, in Adams' sense, because it is used as a quick

⁷Rice (1930) had already criticized that procedure as depending on a select group of people:

Students may be required, good natured academicians may be cajoled, and sundry needy persons may be paid to sort cards containing propositions into eleven piles. But it is difficult to imagine securing comparable judgments, or satisfactory measurements in the final application, from bricklayers, business men, *Italian-Americans*, nuns, stevedores, or seamstresses. (p. 190, my emphasis)

The melting pot still had some lumps.

⁸Item Response Theory (e.g., de Ayala, 2009), the pinnacle of modern scaling theory, dispenses with Thurstone's procedure of readers' sorting items into piles by estimating item locations from the data at the same time it estimates person locations. Some of the necessary mathematical tools, and the computers to implement them, were not developed until recently (Bock, 1997).

method of identifying mineral specimens in the field, by means of a series of scratch tests. It would be intrinsic if it were used primarily to predict which rocks scratched others. Adams' point is that we are not really interested in the properties of the Mohs scale as a *representation* of "real" hardness so much as we are interested in it as a rough *indicator* of hardness. If we do not conceive of our scale as a representation of some underlying dimension, then there is really no empirical meaning to be "lost" by various arithmetic transformations; what we end up with is as good as what we started with. In psychology, it seems clear that, outside of psychophysics, virtually all our measurement is of the extrinsic variety and could be understood better in the terms of an informational theory, like Adams'. When we construct an anxiety scale, for instance, by making up a questionnaire and summing people's responses to get their scores, it is generally because we are interested in the relation of anxiety to some other characteristic; and here, as with the Mohs index of hardness, we need our scale primarily as an indicant of anxiety. Hence, from Adams' perspective, we need worry about permissible statistics for various scale types only in the very restricted area of intrinsic measurement. Otherwise, it is not transformational invariance which determines meaningfulness on a scale, but more nearly the reverse. Stevens would say that differences between hardness numbers on the Mohs scale are not meaningful since the scale is ordinal; but Adams suggests that if such differences were found to correlate highly with some other variable, then this would give them meaning. Rozeboom (1966) makes a similar point:

The reason *why* the IQ scale has no fixed—or rather, meaningfully fixable—zero point is that intelligence has no known feature which can be represented by a numerical scale property whose transformational invariance requires a fixed zero. That is, if we knew how to interpret "IQ score x is twice as large as IQ score y ," this would give the IQ scale a fixed zero point. On the other hand, if we were to believe that IQ is a ratio scale and inferred from this that "IQ score x is twice as large as IQ score y " is meaningful, then, insomuch as the latter is but a relation between numbers which represent different degrees of intelligence it would still remain to identify the binary intelligence relation which holds for any two intelligence levels when the corresponding IQ numbers stand in ratio 2:1. In this way, the [Stevens] perspective disastrously inverts the proper order of inquiry by seeking to answer questions about scale-property meaningfulness in terms of the scale's type rather than by judging a scale's type in terms of what on it is meaningful. (p. 196).

The measurement question in psychology was disposed of in much the same way as other questions of meaning. Baker, Hardyck, and Petronovich (1966) did some empirical simulation research showing that the t test was minimally distorted by nonlinear scale transformations, and most psychologists heaved a sigh of relief and went on with their parametric statistics. A few diehard followers of Stevens (1951) and Siegel (1956) continued to protest parametric analyses of psychological research data, but they have had a hard time prevailing against the sheer fact that parametric analyses can still be performed and will still give answers.

On the other hand, it is hardly to the credit of psychology that essentially all of its measures are extrinsic, in Adams' sense. Loevinger (1957), in her monograph "Objective Tests as Instruments of Psychological Theory," deplored the fact that the library of psychological measures were developed for practical ends, e.g., as clinical screening tools, and as such were utterly unsuited to research in psychological

theory. Though Adams' terminology was not yet available, it was intrinsic measures, as representations of putative psychological quantities, that were needed for that purpose.

The solution of the practical and theoretical problems of measurement had far-reaching consequences for the discipline of psychology as a whole. Among those which Hornstein (1988) lists are (a) the resolution of boundary problems in psychology, giving it a means of distinguishing itself from philosophy on the one hand, and what subsequently became known as pseudoscientific disciplines (e.g., palmistry, phrenology) on the other; and (b) the unification of the field into what Koch (1976) would call a single “community of discourse” with uniform criteria for what is meaningful or scientific. The history of statistics in psychology and the history of psychology itself, thus, parallel the history of the institutionalization of the quantitative perspective.

The entire field was opened up to statistics by the pervasion of measurement, but until that universal quantification was established, statistical methods were largely limited to the domain of mental measurements. With respect not only to statistical methodology but also cultural values, the psychometricians worked in very much the same space as the Pearsonian biometricians. For economic and political reasons, however, they enjoyed a much more spectacular success. The first two decades of the twentieth century were notable for a process that could crudely be described as the application of assembly line methods to American public education and to the military, which offered particular promise for the implementation of bureaucratic, or “scientific,” management principles.

American public education in the first decades of the twentieth century faced severe challenges. The major one was the recent huge wave of immigration, which greatly increased the number of students to be accommodated, but also their diversity, including wide variation in English proficiency. Government schools were supposed to provide a uniform education for all, but curriculum planners were subject to intense pulls in opposite directions. Conflicts over sex education wouldn’t arise for another half century, but evolution was already a controversial topic, as was the conflict between more traditional and vocational training. Business interests, well represented on school boards, often valued ability with machines over fluency in Latin or Greek.

Into these tensions stepped Frederick Taylor with his *Principles of Scientific Management* in 1911. Taylor had studied, with a stopwatch, the motions involved in several different manufacturing jobs, with an eye to determining the maximum efficiency with which they could be performed. His most thorough investigation was of pig iron handlers at Bethlehem Steel. Their job was to carry pigs of iron, weighing 92 pounds, 30 or 40 feet to a railway car, and up a ramp to deposit them on the floor of the car. The 75 pig iron handlers averaged about 12 tons per day apiece. Taylor determined that it should be possible for them to move 47 tons. He selected one man, whom he called Schmidt, and persuaded him to move 47 tons a day for a 60% wage increase. It evidently helped that, in Taylor’s appraisal, Schmidt was “so stupid and so phlegmatic that he more nearly resembles in his mental make-up the ox than any other type” (quoted in Callahan, 1962, p. 40). Taylor and his acolytes,

boosted tremendously by the popular press (Callahan, 1962), made much of the economic gains to be made from application of his “principles of scientific management,” so long as there were plenty of people, dumb as an ox, who were willing to do 300% more work for 60% more pay, and survived to enjoy their new riches.

A not inconsiderable stretch was required to translate Taylor’s system to the field of education. Despite its being the heyday of Watsonian behaviorism, at least a few teachers were baffled by how to analyze the teaching of, say, music appreciation into its component motor movements. But the press for “scientific management” of schools—who could oppose the concept?—was powerful. Outrageous metaphors were taken with utter seriousness: School buildings were now “the plant”; pupils were explicitly referred to as “the product”; and John Franklin Bobbitt, of the University of Chicago (who had received his Ph.D. at Clark University under G. Stanley Hall) asserted that “education is a shaping process as much as the manufacturing of steel rails” (quoted in Callahan, 1962, p. 81).

Remarkably, no parties to the original debate, nor Callahan (1962) in his review, ever noticed the inapplicability of market efficiency to a monopolistic, bureaucratic operation like education, where there is no pricing mechanism. Dissatisfied customers, unless they are wealthy enough to pay for private school tuition in addition to taxes for public schools, can’t simply take their business elsewhere; they have to try to elect a new school board; if they are ultimately successful, it will seldom be before their children have graduated. Schools which are doing a bad job don’t go out of business; they get voted a higher budget. Consider, for example, the US Postal Service, where first-class rates went up 1000% between 1950 and 2000. A private, competitive postal business, which saw its costs going up 1000% over this period, would long since have gone out of business; Congress simply appropriated more money, and could raise postal rates at will, given the legal monopoly.

There were several groups who profited from the Taylorism craze. One was the large number of people offering consultation services in Taylor’s methods and their application to education. Another was the profession of educational administrators, which really came into existence in this period, when the first graduate degrees, with an emphasis on finance, rather than education protocols or philosophy, were granted, by institutions like Columbia. And a third was psychologists.

The new class of educational administrators needed research in order to justify their organizational ambitions from a scientific point of view, although “not infrequently,” according to Danziger (1984), “administrators simply needed research for public relations purposes, to justify practices and decisions which they judged to be expedient” (p. 2). In psychometrics, particularly in the multiple-choice test, they found an important tool. Samelson (1984) credits Franklin J. Kelly with inventing this device in 1914/1915, in the Kansas Silent Reading Test. It was quickly picked up by Otis for use in intelligence testing; he and Terman joined Yerkes’ group in producing the Army Alpha intelligence test in 1917. (The Beta version was a non-verbal test for illiterates.) By 1921, Samelson says, nearly 2,000,000 soldiers and over 3,000,000 schoolchildren had been tested with the new method.

Small—circa 6,000 officers and 200,000 soldiers even in March 1917—led by an officer corps trained primarily at West Point, and composed of career soldiers, the American army of the prewar years was a relatively intimate, almost small-town organization in which soldiers could be placed and rated on the basis of long-term familiarity and a well-entrenched military culture could easily be maintained and propagated. (Carson, 1993, p. 282)

With the Congressional declaration of war, at Wilson's urging, on April 6, 1917, the military suddenly confronted the problem of processing enormous numbers of recruits. Under the influence of the Progressive movement, Carson suggests, the military was predisposed toward solutions that could claim to be "scientific." Psychologists wasted no time: The same day that war was declared, a group of 40 of them met at Harvard to consider the opportunities. Within the month, Robert Yerkes proposed a program of intelligence testing for recruits. The military had never particularly thought of intelligence as a requirement for a good soldier, but it was what psychologists thought they were good at measuring. In short order, his modest initial proposal for testing recruits identified as questionable was transformed into a program for testing all recruits. The administrators—the top echelon—were an easier sell than the commanding officers, who tended to think that basic training afforded sufficient opportunity, as it always had, for observing new recruits. Those officers who were most supportive cited the correspondence between test results and their own ratings of the soldiers—but the officers' ratings had been a major validating criterion of the test, so the test was less of an independent validation of their astuteness than they may have taken it to be. The Navy was also a tougher sell than the Army, out of concern about the ease of malingering—faking bad to avoid conscription. As typically happens with formal, mechanical administrative devices, tests or regulations, the Army Alpha was frequently used in practice to justify decisions made on other, more informal, personal grounds (Carson, 1993).

The recent large wave of immigration had two effects which are relevant in this context. One was a perceived threat to employment. As Patrick Brookhouser (1984) and many others have noted, it was at this time that the concept of adolescence was invented, so to speak, to mark a postponement in entering the work force. It was also the period when nearly all the modern youth organizations were founded—the Boy and Girl Scouts, the Boys' Club, YMCA and YWCA, 4-H, Future Farmers of America, Future Veterans of Foreign Wars, and so on. It is possible, in other words, that the social changes engendered by mass immigration had something to do with the emergence of this new class of social and educational engineers; and the creation, in adolescence, of a whole new social constituency, with its own needs and concerns, was one of its principal achievements.

The second respect in which the new immigrants were perceived as a threat was adulteration of the American gene pool. There was widespread concern that foreign nations were sending their least desirable stock and that traits like alcoholism, sexual deviance, and feeble-mindedness were inherited. The eugenics movement remained active in Britain under the leadership of Galton and then Pearson and Fisher, where perceived causes and remedies had only to do with differential birth rates between social classes. In the United States, the eugenics movement helped to

lead to the passage of the Immigration and Naturalization Act in 1924; the quotas it established, on the basis of ethnic representation several decades earlier, remained in effect until 1965. The task of documenting the intellectual inferiority of different ethnic groups also provided a major stimulus for the growth of psychological testing. (For a detailed history of the eugenics movement, see Kevles, 1985.)

Significantly, psychological measurement was developed, from beginning to end, not to meet any perceived need, for it or for its products, but for the benefit of the emerging profession of psychologists.

9.2 From Large-Sample to Small-Sample Theory in Psychology

For the present purposes, the cultural context of the testing movement is relevant primarily in having given wide prominence to a statistical methodology and its application to psychological phenomena. From a statistical point of view, the work was much the same in psychometrics as in biometrics; the interest in both areas lay in aggregate characteristics of biologically or socially defined populations. As was noted in Chap. 7, the distinction between sample and population was not so salient as it was soon to become, owing simply to the very large size of typical samples. The statistics texts for psychologists in that period, on through the 1930s, were wholly given to large-sample theory, with a function that was essentially descriptive. The widely used texts of Kelley (1923) and Chambers (1925) contain no mention of inferential statistics and the small-sample theory which were to become the mainstream in psychological statistics. Thus it was that Boring, writing in 1920, could speak of a sample of over 300 as “very small” (p. 24). As the domain of application of these methods spread, of course, there was a natural tendency for sample size to decrease from thousands to hundreds to dozens, creating a certain readiness for the small-sample methods to follow.

From the point of view of psychological research, on the other hand, psychometrics transformed the field. Psychological research had been conducted up to that time under what Danziger (1987) labels the Wundtian paradigm, where the object of inquiry was processes in individual minds. Typical experiments used only a single “subject” (a term taken from nineteenth-century medical practice), often the experimenter; additional subjects were used literally as replications. Psychometric research exemplified the Galtonian paradigm, as Danziger calls it, where the interest was in aggregates rather than individuals, and the concept of error pertained, as with Quetelet, to deviations of individuals from the norm rather than to limitations of the measurement process.

This shift from Wundtian to Galtonian methods left, however, a large gap between psychological theory and research. But for the timely advent of Fisherian small-sample methods, it is possible that psychological research would have remained more Wundtian, with psychometrics aligned with a more sharply differentiated field

of educational research. Danziger explains how small-sample theory allowed that gap to be closed instead, through the concept of the treatment group. Though Danziger downplays Fisher's influence, the source of the metaphors was obviously agricultural, as the language of experimental design continues to show; Lovie (1979) remarks that early expositions of analysis of variance were too "redolent of the farm" (p. 167) for psychologists easily to understand. The hybrid constituted by the experimental treatment group Danziger calls neo-Galtonian, since the statistical treatment was Galtonian. The application of different treatments or the imposition of conditions under the experimenter's control, however, allowed inferences to be made about psychological processes, albeit in the aggregate rather than the individual. It was thus also necessary for psychological theory to follow suit, in redefining its object as statements about characteristics of aggregates.

In this it was scaffolded by the ascendant positivist philosophy of science. In disallowing reference to what could not be observed by several people at once, logical positivism had the well-known effect, to the extent that it was accepted, of making psychology behaviorist. No less important from the standpoint of methodology, however, was exclusion of reference to causality. The exclusion of causality from science removed any possible basis for a distinction between the general and the aggregate; it left only coincidental aggregates, as Lonergan (1970) calls them, as the object of scientific knowledge (Chap. 10). With nothing to be grasped but statistical regularities, the emphasis was correspondingly heavy on prediction and control, as stand-ins for understanding.

The prevailing philosophy of the early twentieth century could thus be characterized crudely as a bold, formalized acceptance of the crisis in knowledge bequeathed by Hume in the eighteenth century. Its view of induction, innocent of causality, accorded perfectly with that instantiated in Laplace's Rule of Succession. The fascination of philosophers for a century, the Rule was never used by scientists (including Laplace himself, in his major contributions to astronomy), with obvious reason. Scientists in other fields went about their work on the basis of an informal, everyday understanding of epistemological concepts, occasionally, as in physics, venturing a reworking of them to accommodate anomalous observations. They never encountered a need for special rules to tell them how to draw conclusions from data.

Statistical inference entered scientific work for the first time in twentieth-century psychology, to shore up, it would appear, a methodology that needed the support. The Law of Succession itself, had nothing more appealing come along to supplant it, would not really have filled the need. Apart from the fact that it would have been rejected as dipping, with the principle of indifference, into metaphysics, it did nothing to address the issue of small-sample research on treatment groups and was limited in any case to binary characters.

It is ironic that the work of Fisher, who vehemently rejected the arid positivist doctrines (in favor of what, it is harder to say), should have been enlisted in support of such a methodology. The general process was described in Chap. 7. Laboratory research in psychology, like agricultural field trials, yielded small quantities of data. Following established large-sample procedures, experimental psychologists conceived their object as making statements about the population from which their

sample was drawn. Only, as Gosset had observed (Student, 1908), samples of 30 observations or so are largely useless for the purpose of estimating population parameters. The variability of the estimates is simply too great. Fisher's contribution, intentional or not, was to make a silk purse out of a sow's ear: Estimation was largely abandoned in small-sample theory for significance testing, which yielded, in place of an estimate of the magnitude of a treatment effect, a yes-no answer to the question of whether the treatment had an effect or not. For practical application, such an answer would have no real value, but significance testing was enlisted primarily in the service of building theories with only remote implications for application, and in this role it performed beautifully. It always gave an unequivocal answer to research questions about whether there were a difference between groups.

The concepts of the treatment group and of significance testing still left some gaps in the methodology, two of which should be noticed here. In the first place, the appeal of statistics is intrinsically to large numbers. The concept of large numbers itself poses problems that were noticed in Chap. 6, but in the present context, the point is that statistical inference might have been expected to drive psychological research methodology back to large samples after all. The other is that the population, ostensibly the object of interest, has tended to fade from view. The two issues are related. Indefinitely large numbers are necessary at some point for statistical inference. In significance testing, they reside both in the population, presumed infinite, and in the sampling distribution, also infinite for parametric tests. But both of these, in ordinary research applications, are entirely hypothetical. The only thing that is ever observed is a small quantity of data; the context of inference is what might have happened in a hypothetical process of repetitive random sampling from a hypothetical infinite population. And, in truth, for practicing experimental psychologists the population and the sampling distribution were philosophical niceties in which they had little interest. The new methodology took hold just because it allowed a certain kind of research practice to proceed.

In broad strokes, then, the development of contemporary psychological research practice proceeded through the following steps: (a) the first “empirical” psychological research by Wundt and his followers, on mental processes of individuals; (b) the (largely unrelated) emergence of the mental testing movement, using the large-sample Galtonian statistical methods of the biometrists; (c) the adoption from agricultural research of the concept of the treatment group, combining the Wundtian and Galtonian approaches; (d) the concomitant introduction of small-sample statistical analysis from agriculture; and (e) the redefinition of the goals of psychological research to fit the new methodology.⁹ The consequences of the transformation were extremely far-reaching, as Danziger (1987) has discussed, particularly in the prescriptions it established for “paradigmatic” research—that which would be supported by the profession and, increasingly, by the government.

⁹For the incorporation of the specific technique of analysis of variance into psychological research, see Lovie (1979, 1981) and Rucci and Tweney (1980).

It is instructive to consider the speed with which the transformation took place. From the viewpoint of Fisher's followers, the acceptance of his methods may well have seemed slow. It was 10 years after the publication of *Statistical Methods for Research Workers* that the next textbook of Fisherian methods was published—just as the *t* test, although known, had not been used before Fisher. The availability of small-sample methods did not suffice to create a market for them. Even as the need and justification for them were developing together, the change in patterns of thinking and use was more gradual than abrupt. Danziger (1984), surveying research articles from 1914 to 1936, found that the Galtonian paradigm, oriented to group data, predominated in the *Journal of Educational Psychology* and the *Journal of Applied Psychology* from the beginning, but only gradually gained acceptance in the *Journal of Experimental Psychology* and the *American Journal of Psychology*. The former are obviously distinguished as applied research journals, but Danziger (1987) notes significantly that:

When one speaks of applied psychology in this context one must be clear about the fact that a very specific kind of application of psychological knowledge is involved. It is an application that takes place in an administrative context, and the individuals to whom such knowledge is useful are administrators. The kind of research that is involved here is not designed to enhance the self-understanding of individuals, but to provide administrators with a data base from which to make more effective decisions about the programs and individuals under their control. The psychological studies initiated in this context were generally characterized by the following features: (1) the goals of the research were determined by issues of specific and immediate social concern rather than by issues of general psychological theory; (2) the questions asked were questions of output, performance and efficiency, rather than questions involving internal psychological processes; (3) the subjects on whom this research was carried out, and to whom its results applied were assumed to be *minors*, either in legal fact or in the more general sense that they were persons without valid insight who were not free to determine their own fate but were objects of social control by those in authority. (p. 43)

The social status distinction between experimenters and subjects, virtually necessitated by behaviorism, has persisted to this day (see Schultz, 1969); only in recent years have alternative models been put forward in mainstream psychology (e.g., Forward, Canter, & Kirsch, 1976).

When small-sample methods were introduced into psychological statistics texts, their presentation clearly showed the influence of psychometrics: Significance tests were discussed under the heading of “reliability.” Thus Sorenson’s (1936) chapter on the subject is “The Reliability of the Difference between Means”; Cooke’s (1936) is entitled “Measures of Reliability.” Though these books postdate the first edition of *Statistical Methods for Research Workers* by fully a decade, they both use as the rejection criterion three or four times the probable error, which Fisher denigrated¹⁰; there is no mention of probabilities of .05 and .01.

¹⁰As possible testimony to Fisher’s influence, it is almost amusing to notice the vehemence with which McNemar (1949) put down the use of the probable error, as a “needless and antiquated” “nuisance practice” (p. 89). Even mild affect stands out in most statistics texts, and it seems espe-

In most texts prior to World War II, small-sample theory remained a specialized topic. Even Yule (1911), in his pathbreaking text, mentioned Student in the reference list but did not discuss his work; the *t* test was not discussed until the 11th edition (1937), which was revised in collaboration with M. G. Kendall. Cooke, in his *Minimum Essentials of Statistics*, remarks that:

If N is less than 30, there is little justification for using the measures of reliability given in this chapter. There are formulas available for determining the reliability of measures based on a small number of cases, but such formulas are not in common use and at best are strictly applicable to experimental situation with which the student in beginning educational statistics has had but little contact. (1936, p. 105)

Croxton and Cowden, in their *Applied General Statistics* (1939), presented a section on small-sample methods, but starred it as optional.

The first textbooks to carry the Fisherian message appeared in the years just before World War II. The first and most successful of these was Snedecor's *Statistical Methods*, the first edition of which was published in 1937. Snedecor had long been an admirer of Fisher's and, as was noted in Chap. 7, had invited him to Iowa State on two occasions in the 1930s. His book was the first one after Fisher's explicitly to accord primacy to small-sample theory. The decade following World War II saw an enormous number of statistical manuals into print, but the situation by this time was complicated by the fact that the Neyman-Pearson theory was also becoming known. The antagonism between these two camps—the intensity of the debate coupled with the subtlety of the issues—posed a profound problem for authors of practical manuals; the effects are felt down to the present day. Conscientious authors of advanced texts—Johnson's *Statistical Methods in Research* (1949) and Anderson and Bancroft's *Statistical Theory in Research* (1952) are excellent examples—made a point of studying and presenting both theories, though not necessarily as antagonists. Generally the procedure in these texts was to use Neyman-Pearson theory to supplement the Fisherian rationale, especially with respect to the concepts of power, alternative hypotheses, and Type I and Type II errors. Authors of more elementary texts then seem most often to have used these works as their primary sources in turn; with each generation, down to the ultimate consumer, oversimplification and distortion increased. We may observe the process in connection with the concepts of probability, significance versus confidence, power, and random sampling.

9.2.1 The Concept of Probability

The fundamental problem is the concept of probability and the implications of the frequency theory for statistical inference. Nothing at all in a strict frequency theory warrants the use of significance levels in an evidential manner; Neyman always

cially odd when directed to a procedure that is logically equivalent to the one that is advocated in its stead.

insisted that a decision to accept or reject has nothing to do with whether we *believe* it. But virtually all textbook authors have followed Fisher on this point, rather than Neyman and Pearson, even when they managed a conscientious exposition of significance levels and confidence coefficients in terms of direct probabilities. The statement of Edwards (1950) is typical: “A probability of .01 or smaller will be regarded as *very significant* and will simply mean that the hypothesis being tested will be rejected with *greater confidence* than when the probability is between .05 and .01” (p. 28; his emphasis). The characterization of probabilities between .05 and .01 as a “region of doubt” (e.g., Johnson, 1949—who may merely have been following the precedent of Neyman and Pearson themselves, 1933b) is further indication of an underlying epistemic attitude toward probability.

A few authors discussed the concept of probability explicitly; those who did, interestingly, seemed often to take the interpretation of Boole and Reichenbach (though without mentioning their names), that probability refers to statements rather than events (e.g., Clark, 1953; Mills, 1955). And others, without much explicit discussion of probability, still got the account right when they came to present the theory of testing and estimation. Sorensen (1936), Snedecor (1946), McNemar (1949), Edwards (1950), Anderson and Bancroft (1952), and Kempthorne (1952) are among those who correctly characterized significance levels in terms of forward probabilities. The last of these was especially clear on the distinction between forward and inverse probability; he said forcefully: “It should be noted that the level of significance has no relation to the probability of the hypothesis being true, and, in fact, no such probability exists” (Kempthorne, 1952, p. 12).

Walker and Lev (1953), on the other hand, while they were generally careful and thoughtful, went so far toward simplifying the concepts of statistical inference for beginning students that their description of confidence intervals sounds misleadingly like inverse probability: “The concept of probability is used in reasoning from a known population to a random sample. The concept of confidence is used in reasoning from an observed sample to its unknown population” (p. 56).

Quite possibly the worst offender with respect to the distinction between forward and inverse probability was Guilford, whose *Fundamental Statistics in Psychology and Education* was also one of the most successful of such texts. The first edition appeared in 1942; the sixth edition, in 1978, was coauthored by Benjamin Fruchter. Guilford (1942) announced his confusion very boldly in the headings in his chapter on “Testing Hypotheses”: They include “Direct Determination of the Probable Validity of a Null Hypothesis” and “Probability of Hypotheses Estimated from the Normal Curve.” When he spoke of the area under the tail of the normal curve as indicating the probability of a true difference between means in a given direction (p. 167), it is clear that he has implicitly shifted, with Fisher, from thinking of the normal curve as a distribution of sample mean differences to thinking of it as a distribution of the true mean difference. And the way he spoke of the “odds against the null hypothesis” (e.g., p. 164) would gratify a contemporary Bayesian. Some changes, of course, were made over the years. By the third edition in 1956, he had ceased to present significance testing under the heading of reliability, and by the sixth edition of 1978, he had implicitly corrected the quasi-fiducial interpretation

just described. But the worst confusions persisted right on through successive editions: e.g., “*We cannot prove the truth of the null hypothesis; we can only demonstrate its improbability*” (Guilford 1942, p. 151; emphasis in original). But Guilford is scarcely alone. Compare the statement in another widely used modern text:

Scientists want to draw inferences from data. Statistics, via its power to reduce data to manageable forms (statistics) and its power to study and analyze variances, enables scientists to attach probability estimates to the inferences they draw from data. Statistics says, in effect, “The inference you have drawn is correct at such-and-such a level of significance. You may act as though your hypothesis were true, remembering that there is such-and-such a probability that it is untrue.” (Kerlinger, 1973, pp. 186–187)¹¹

If, in turning significance levels around to the exact opposite of their meaning in Neyman-Pearson theory, Guilford and Kerlinger intend an outright rejection of that theory, they are unusual among contemporary writers of textbooks in psychological statistics; one suspects that, as William Cuffe said of Cobden, if Neyman were alive, he would roll over in his grave.

9.2.2 Significance Versus Confidence

Very often estimation and hypothesis testing were confused. One reason, no doubt, was the fact that within Neyman-Pearson theory the calculations for confidence coefficients and significance levels are the same. Another was the failure to grasp the significance of the shift from large-sample estimation methods to small-sample hypothesis testing. Clark’s *An Introduction to Statistics* (1953), Adams’ *Basic Statistical Concepts* (1955), and Guilford and Fruchter’s *Fundamental Statistics in Psychology and Education* (1942) are among those which speak of “confidence levels” in connection with tests of hypotheses. Adams’ text, a relatively mathematical treatment, did not use the term “significance” at all.

Estimation was typically presented in terms of confidence intervals rather than fiducial probability, even if the interpretation given, as was often the case, was closer to the latter. The occasional presentation of fiducial probability (as in Croxton & Cowden, 1939, and Snedecor, 1946) hewed close to Fisher’s own exposition; more commonly, as in McNemar (1949), fiducial probability was mentioned only in passing and somewhat mysteriously (McNemar, for one, deleted it altogether from subsequent editions).

In general, significance testing was emphasized to the neglect of estimation. A few authors (e.g., Churchman, 1951) protested the ritualistic use of .05 and .01 levels, and several of the more advanced texts (e.g., Cochran & Cox, 1950; Kempthorne,

¹¹ Kerlinger errs not just on issues of interpretation but in the statement of mathematical theorems. After saying that the law of large numbers is so simple one wonders why it took Bernoulli 20 years to work it out, he gives a statement of it which is patently false:

Roughly, the law says that as you increase the size of samples, you also decrease the probability that the observed value of an event, A , will deviate from the “true” value of A by no more than a fixed amount, k . (1973, p. 190)

1952) maintained that estimates of the magnitude of effects were more appropriate an object of inquiry than significance levels. But the principals of the theory had already provided too generous a warrant for significance testing, in spite of their own occasional caveats; confidence intervals and descriptive indices of association failed to take hold in the research literature.

9.2.3 Power

The concept of power was unfamiliar to the first generation of psychologists trained on the Fisherian manuals up to the early 1940s. It was slow to catch on, and was often regarded as an “advanced” topic. (The difficulty of power calculations afforded a certain practical justification for this policy.) The advanced texts themselves appear to have had some trouble with the concept; Johnson, in his *Statistical Methods in Research* (1949), spoke in one place (p. 64) of β , the probability of a Type II error, as the power of a test; two pages later he got it right, giving $1 - \beta$ as the power. Anderson and Bancroft (1952) defined β as power and used it that way consistently, if idiosyncratically. Guilford, however, continued to dismiss power as “too complicated to discuss” (1942, p. 217), right up until the sixth edition. He may have been right. It is disappointing to see so excellent a text as Cook and Campbell’s (1979) *Quasi-Experimentation* consistently confusing power and significance—more precisely, $1 - \beta$ with $1 - \alpha$ —as in speaking of an “effect that could have been detected in a particular experiment with, say, 95% confidence” (p. 46; cf. p. 40).

Until recently, most texts, whether or not they were as glib about it as Guilford, still gave slight attention to power. A few protested: Both Kempthorne (1952), in his original text, and Anderson and Bancroft (1952) made the point that if significance is the only criterion, no statistics are necessary; a random device, such as rolling a regular icosahedron, will yield an α of the desired level. But only in the last few years has the concept of power received much attention in the textbooks. The increase in emphasis between the first edition of Hays’ *Statistics for Psychologists* in 1963 and the second in 1973 is indicative of the trend. And beginning with Welkowitz, Ewen, and Cohen (1976), it has become more common for introductory texts to contain a whole chapter on power analysis.

9.2.4 Random Sampling

Many of the early statistics texts contained extensive discussions of the concepts of random and stratified sampling, which were appropriate especially for educational, psychological, and sociological survey research; but emphasis on these concepts in general psychological statistics texts declined over the years, until stratified and cluster sampling reappeared as a special topic in its own right. The examples of random sampling given in the older texts—opinion polling, or Snedecor’s (1946)

sampling of hog farmers in eastern South Dakota—had little relevance to laboratory research in psychology; and the actual procedures used to obtain subjects in psychological research made detailed discussions random sampling a little embarrassing, if not altogether unnecessary. Fisher's interpretation of the process of statistical inference, though rarely called upon for an explicit defense, provided nonetheless a tacit justification: For in his theory the investigator starts logically as well as factually from the data at hand and extrapolates backward to a suitably conceived hypothetical population.

Confusion persisted, however, over the concepts of randomness and representativeness, as small-sample methods were assimilated to large-sample theory, which was sometimes not fully understood in the first place. The upshot was that small-sample inference was often justified on the grounds of representativeness—and the results are obvious today in the findings of Tversky and Kahneman (1971); see Chap. 10. Among statistics texts, possibly the worst offender in this respect was Smith's *Elementary Statistics* (1934), which promulgated a total inversion of the two methods. After making the surprising statement that “Chance plays no part in modern statistical theory” (p. 298), he went on to articulate two rules of statistical inference:

The first rule is that the portion of the total population being studied for the purpose of making generalizations about the whole, must be similar to the whole in the characteristic about which the generalization is to be made (p. 324);

the second rule of statistical inference is that the data must be homogeneous (p. 325).

Thus the first rule stipulates representativeness, and the second requires that variability, which is the stuff of statistical inference, be absent. “According to the theory of probability, if a sample group of data has been selected at random it *will* be representative” (p. 325). If it were representative, of course, the inference would be straightforward and deductive, with no need for statistics.

9.3 To the Present

The authors of the statistics manuals are not entirely to blame; the equivocations and inconsistencies in the original documents made a clear presentation to elementary students a rather hopeless task. And if the manuals could be grievously faulted on all the counts I have just discussed, uncontestedly they also did something extremely well. They delivered to an eagerly waiting public a powerful appearing new tool, whose promise was rivaled only by that of electronic computers. It was a seller's market. Snedecor's *Statistical Methods* was the first intelligible manual to present the Fisherian methods; in spite of being published by a small college press, it was a huge success, and is still widely used; a quarter of a century after its publication, it was the most cited book in scientific work, according to the *Science Citation Index* (Kempthorne, 1972).

To achieve their success, the manuals aimed ultimately to present the theory and techniques of statistical inference in so simplified and codified a form as to make no more demands than a recipe in a cookbook. Writing a cookbook of statistics so as to spare the user any serious thought entailed, in the first place, that mathematical and philosophical rationales be dispensed with; only the end product was presented.¹² The rationales were all open to question, and the subject of heated debates among their inventors. A cautious, considered treatment would not “sell,” any more than a vacillating psychotherapist or an honest politician. Theoretical disputes were minimized, as embarrassing in-fighting, and the edifice of statistical technique was made to look monolithic and cohesive. Subtleties were avoided. As a corollary, all ideas were presented anonymously, rather than as the invention of particular theorists. Fisher is known because he wrote one of the cookbooks himself—his *Statistical Methods for Research Workers* was for over a decade the only source on his methods available—but the names of Jerzy Neyman and Egon S. Pearson are scarcely household words among psychologists. The presentation of ideas as authorless has the further great consequence that they look just like the truth, instead of somebody’s particular theory. The theory was thus made all the harder to question. And it was mathematics, after all.

The big dilemma was the issue dividing Fisher and Neyman and Pearson. They claimed the same starting point—the frequency theory of probability—but ended up in different places. Fisher wanted a theory of inductive inference; Neyman and Pearson, after initially situating their theory in that domain, gradually gave up the commitment to inference and advocated a theory of decision making in repetitive sampling contexts. On logical grounds, Neyman and Pearson had decidedly the better theory, but Fisher’s claims were closer to the ostensible needs of psychological research. The upshot is that psychologists have mostly followed Fisher in their thinking and practice: in the use of the hypothetical infinite population to justify probabilistic statements about a single data set, in treating the significance level evidentially, in setting it after the experiment is performed, in never accepting the null hypothesis, in disregarding power. Fiducial probability is the one point on

¹²In some respects what was happening in statistics books reflected larger changes in American life during the same period. If we look at actual cookbooks, we can see some of the same trends. Rombauer’s *The Joy of Cooking* (1931) first appeared in the same decade as Snedecor’s *Statistical Methods*; compared with the latest edition, the first contained much fuller discussions of the whole process of food preparation (e.g., how to dress a rabbit), which cooks of the time commonly involved themselves in. When I consulted my 1943 edition of *The Joy of Cooking* on how to roast a turkey, the first instruction was: “Draw a turkey.” (It is interesting how quickly we have lost the language for things we no longer do. One imagines young people today being thoroughly confused by such instructions.) Nowadays everything is already so thoroughly processed by the time it reaches us that we have largely lost the sense of working with animals and plants when we are cooking, just as we have lost the sense of working with human beings when we are analyzing experiments. We reach for a multivariate ANOVA program as casually as we reach for a frozen lobster newburg.

Users of cookbooks of any sort would do well to heed Étienne Halphen’s (1955) reminder that good cooking is a fine art, and not a matter of mechanically following recipes, or of picking a canned program off the shelf.

which Fisher lost. But even that loss was in name only: because confidence intervals are almost invariably misinterpreted in such a way as to be equivalent to the discredited fiducial intervals. Yet the rationale for all our statistical methods, insofar as it is presented, is that of Neyman and Pearson, rather than Fisher. The tension between these two authorities permeates the field today. The Neyman-Pearson theory serves very much as the superego of psychological research. It is the rationale given, if still anonymously, by the most sophisticated and respected textbooks in the field, which teach the doctrine of random sampling, specifying significance levels and sample sizes in advance to achieve the desired power, and avoiding probability statements about particular intervals or outcomes. But it is the Fisherian ego which gets things done in the real world of the laboratory, and then is left with vague anxiety and guilt about having violated the rules. The Neyman-Pearson doctrine remains as a persistent source of irritation; its demands seem extraneous and unnecessary to the real business of research; yet mere practitioners cannot begin to muster the authority, or the presumption, to challenge it on its own ground; and the inadequacies of the Fisherian theory—along with the essential arbitrariness of it all—would be exposed at the same time.

The first books and articles entirely devoted to analysis and criticism of the prevailing methodology began to appear in the late 1950s, just as the Bayesian alternative was coming on the psychological scene. Its entrance was thus rendered especially dramatic, as it was cast inevitably in the role of a self-proclaimed savior from the increasingly disturbing problems of the frequentist theory. Bayesian methods had of course been available in some form all along, but had lain in disrepute since the emergence of experimental psychology; Bayesian statistics could be said, as Ebbinghaus said of psychology itself, to have a long past but a short history.

In fact, all work on statistical inference up to the twentieth century could be regarded as essentially Bayesian. Its key feature was the use of inverse probability, via Bayes' Theorem, the Rule of Succession, or some other device. Inverse probability was opposed primarily on the ground that it resisted either a classical or a frequency interpretation, and it declined as the frequency theory of probability rose to predominance. In the twentieth century, several thinkers took a theoretical step which facilitated the revival of inverse inference: the recasting of probability in fundamentally epistemic terms. With that step, direct and inverse probabilities became explicitly all of the same kind. Though the logical and personalist versions of the so-called Bayesian theory of statistical inference differ in the meaning they attach to probability and in the way it is to be known or measured, in practice their solutions to statistical problems are formally similar.

It appeared briefly in the 1970s that Bayesian statistics might be making some actual inroads into psychological research. The first wholly Bayesian elementary statistics text for psychologists was Phillips' *Bayesian Statistics for Social Scientists* (1973). In that same year Hays (1963) included a substantial chapter on Bayesian methods in the second edition of his *Statistics for the Social Sciences*. The following year saw Novick and Jackson's *Statistical Methods for Educational and Psychological Research* (1974), written from a Bayesian point of view; Novick also

wrote a chapter on Bayesian procedures for the second edition of Blommers and Forsyth's *Elementary Statistical Methods in Psychology and Education* (1977).

Interestingly, and predictably to retrospect, the theory in these texts, to the extent that it was presented, was personalist, but the implied practice was that of Jeffreys. The behaviorism of the personalists (at least of de Finetti) is still much more palatable to American psychologists than the apriorism of Jeffreys, but the actual device of betting in a scientific context is not, and the concepts of costs and utilities required further quantification of intangibles. Were Bayesian theory ever to gain ascendancy in American psychology, this is surely the form it would take: lip service to the personalist justification, with practice a pro forma matter of diffuse priors. The illusion that orthodox methods "deliver the goods," as Jeffreys (1939/1961) puts it, is sufficiently strong, however, that the substantial criticisms of them were still not enough to provoke a radical shift to Bayesian methods.

These Bayesian texts were not widely adopted, and Hays' excellent chapter on Bayesian statistics was not often taught by instructors using that book. The texts disappeared from print, and Hays dropped the Bayesian chapter from the third edition in 1981 (at the publisher's insistence; Gigerenzer, 2004).

The neglect of Bayesian statistics by American psychologists is partly a matter of timing, but partly a matter of logic. The first detailed Bayesian theory of statistical inference, with a body of technique comparable to the Fisherian corpus, did not become available until 1939, with the first edition of Jeffreys' *Theory of Probability*. By that time, Fisher's methods were well on their way to becoming established. Although Jeffreys and Fisher came out of a similar epistemological tradition (as against Neyman), Jeffreys' explicit rationalism would have prevented his acceptance by American psychology at the time. Had de Finetti's approach become known to English readers in the 1930s rather than the 1960s, it is conceivable that it could have been adopted from the beginning. But—and this is the logical point—contemporary Bayesian methods are really parasitic on Fisherian small-sample techniques and might well also have been perceived as an inessential addition.

9.4 The Context of Use

Two aspects of the way statistical inference is used in psychological research will be useful to consider here.

9.4.1 *Contexts of Discovery Versus Verification*

We can count on the textbooks for a good presentation of the prevailing orthodoxy—not necessarily so much what psychologists actually do as what they say about what they do. Kerlinger (1973) advises us in the quotation above that they

want to evaluate hypotheses probabilistically; Guilford (1942) adds a distinction between the contexts of exploration and testing.

Some experiments are designed very simply to answer questions such as, “if I do this, what will happen?” Such experiments are exploratory. The end result is usually in the form of hypotheses, which need further investigation. A higher type of experiment is one that sets out to test the truth or falsity of some hypothesis. From previous experience, derived from an experiment or not, we suspect that a certain relationship exists, but it requires a crucial test to enable us to accept or reject the hypothesis. If the crucial experiment comes out one way, the hypothesis is probably correct; if it comes out another way, the hypothesis is probably wrong. (1956, p. 203)

Beyond the splendidly forthright Bayesian interpretation, we may also note here the intrusion of moral overtones in the distinction between contexts of discovery and of verification. The distinction between the exploratory, or theory-searching, process and the confirmatory, or theory-testing, aspect of research is often felt to be needed because methods for discovery may not be the same as methods for verification. Verification, in contrast to what the name suggests, is not a process of making true, but of making sure, and hence is not so much something we do to our hypotheses, as something we do to ourselves. It amounts, in one way or another, to looking to see that what we have seen before can be seen again; and a method, once identified, can be used repetitively, until we are satisfied. Discovery, on the other hand, is seeing something for the first time. Rules for discovery are thus hard to come by, because they can only be used once; then they are no longer rules for discovery. Methods for discovery elude specification, moreover, because the requisite attitude is one of receptivity to inspiration; if we followed a rule, we would not call what happened an inspiration, or the result a discovery.

In these terms, the distinction is familiar enough, but it can very easily be overdrawn (Feyerabend, 1975, pp. 165–167). The processes of discovery and verification may best be thought of, perhaps, as marking more or less incidentally two complementary aspects of ongoing inquiry. The phase of verification represents the formalizable aspect of scientific inquiry, and rule-following activity; the process of discovery is unformalizable and represents, as it were, the rule-breaking aspect. Both processes go on in alternation, almost concurrently. To find the distinction actually becoming sharp in practice, Feyerabend suggests, is indicative of a temporary stasis in research; we might go on to say that our tendency to find the sharp distinction plausible is indicative, perhaps, of the static quality of our conception of research.

The relevance of the distinction in the present context, in any event, lies primarily in how seriously we take significance testing. It is—ironically, given Guilford’s status distinction—in exploratory contexts, where we have no special investment in the outcome of a test, that we will be more inclined to accept or reject a hypothesis in just the textbook fashion, without further criticism or rationalization. Where our prior beliefs are weakest, the results of the test will carry the most weight. In the limit, were we conducting the test in a void, we would have reached the condition stipulated by Fisher as necessary for statistical inference: that the sample data contain the whole of the relevant information.

In confirmatory, or theory-testing, inquiry, on the other hand, the significance test is inserted into a more elaborate theoretical network, with rich implications from previous findings. If the results come out in an unexpected direction, or if they fail to reach an expected level of significance, we are reluctant to allow our whole theoretical structure to be overturned on the basis of a single test. And we have plenty of outs, quite apart from the explanation of “chance as the ever-present rival conjecture” (the phrase is Polya’s, 1954). No experiments in psychology, even the ones we ourselves run, are so tight as to debar the possibility of criticism. We always have, and not infrequently exercise, in other words, the option of rejecting our data instead of the hypothesis. That option is excluded from nearly all contemporary theories of statistical inference, though, ironically, as was noted in Chap. 5, significance testing has its roots in the rejection of observations. There are grave risks in the practice, of course; the concept of theory testing can quickly be emptied of meaning if we are always ready to discount offending observations. Data, as their name implies, are supposed to be *given*, and not open to question; but a posture of total subservience to them becomes as indefensible as aloof and utter indifference.

The overall process of theory construction, whether we are looking at its exploratory or confirmatory aspects, is primarily an epistemic enterprise, whatever the place that may be found for “decisions.” As such, it follows the same rules, to the extent that it follows any, as plausible reasoning to any purpose. Consider the matter of discriminative power. A theory receives no special support from experimental evidence unless it successfully predicts results not predicted by other theories, and the degree of support for the new theory increases with the surprisingness of its successful predictions. The extreme example of successful divergent predictions in modern science is probably the theories of Velikovsky (e.g., 1950, 1955). Consider also the matter of resolving discrepancies between theory and observation. Whether we reject the theory on the basis of the observations, or hold onto the theory and reject the observations, depends on the prior support for the theory. This prior support will generally hold some moderate value. If a hypothesis is given no particular credence to start with, it is ordinarily not worth the bother to test; if it is held so strongly that discordant data will have no impact, it cannot be said to be undergoing a test, or even a confirmation, in any honest sense of the word. Here, again, Velikovsky is the clearest example: His theories are given so low a prior probability by many scientists as to make them indifferent to any pattern of support, however striking. Consider two examples from psychology, involving statistical support.

The first is a study by Sapolsky (1964), which was criticized by Lykken (1968). Sapolsky hypothesized that psychiatric patients giving frog responses on the Rorschach would show a higher incidence of eating disorders than patients not giving frog responses. His theoretical rationale was that these patients, the frog responders, held an unconscious belief in the “cloacal theory of birth,” involving oral impregnation and anal parturition, and the frog is a cloacal animal. Sapolsky obtained a highly significant chi-square to support his theory: 19 of 31 frog responders, but only 5 of the 31 controls, had eating disorders. Moreover, there are no obvious alternative psychological theories to explain the high incidence of eating

disorders among frog responders. Lykken, however, flatly disbelieved Sapolsky's hypothesis, even after seeing the results, and his sample of 20 clinical colleagues generally agreed with him.

A second example is the general phenomenon of ESP. Psychologists who are accustomed to interpreting p values as the degree of support for their (nonchance) hypothesis are logically committed to regarding ESP as the most reliably established phenomenon in the field. More common, however, is a response like that of Hansel (1966), who argued that the incredibly small p values reported in ESP research prove that the experimenters must be cheating. The ESP example is no longer hypothetical, thanks to Daryl Bem (2011) having published a series of experiments in the *Journal of Personality and Social Psychology*. Bem's procedures, and analyses, were fairly careful; he did use one-tailed tests, but they were perhaps as justified there as anywhere. His overall effect size of .22, he observed, was comparable to effect sizes in other social psychological research. In an article entitled "Why Psychologists Must Change the Way They Analyze Their Data," Wagenmakers, Wetzels, Borsboom, and van der Maas (2011) sharply criticized him for not having done a Bayesian analysis. They emphasized that their failure to replicate Bem's results, in reanalyzing his data, was *not* due to their use of a prior making the alternative hypothesis incredible; but, somewhat disingenuously, they stipulated an infeasibly large effect size for H_1 , to make the Bayes factor (Jeffreys' K) favor the null. As Bem, Utts, and Johnson (2011) pointed out in reply, if ESP effects were that large, ESP wouldn't be controversial. A more adequate response was made by Galak, LeBoeuf, Nelson, and Simmons (2012), who simply replicated Bem's research, using his procedures and analysis—and got null results.

In general, the problem of the equivocal status of significance tests arises partly because there are always "gaps" in the inferential chain, between scientific theories and their translation into statistical hypotheses. The links are never so tight but that an alternative hypothesis can be squeezed in. As a last resort, the cheating hypothesis is always available; like Descartes' demon, it cannot be disproved. The consequence is that significance tests are meaningful as a criterion for deciding between theories only in that community of scientists who are prepared to accept the experiment as intrinsically valid. To others, who are prepared to criticize the experiment, or otherwise to reject its results, the test cannot speak. Accordingly the power of significance tests as arbiter of theoretical disputes is reduced; the really important theoretical boundaries cannot be crossed by this supposedly objective instrument, and the impact of experimentation is limited to the comparatively minor disputes arising within a given framework. The plethora of theoretical perspectives in contemporary psychology makes it apparent that these "communities of discourse" (Koch, 1976), sharing a common framework for the arbitration of disputes, will often be rather small.

9.4.2 Statistical Significance as an Indicator of Research Quality

In recent years, a number of writers (e.g., Atkinson, Furlong, & Wampold, 1982; Greenwald, 1975; Sterling, 1959; Walster & Cleary, 1970) have drawn attention to the fact that we treat statistical significance as a desirable object in itself and as an indication of the quality of research. Sterling (1959) and Atkinson et al. (1982) have documented that journals overwhelmingly publish articles reporting statistically significant results. Atkinson et al. submitted a manuscript, less the discussion section, to 50 consulting editors, and found that the same piece of research was rather clearly accepted or rejected according to whether the results were significant or not.¹³ And certainly many graduate students can testify to having been sent back to collect more data for their dissertations or to revise their hypotheses when the results were not significant.

Logically one would think that the merit and informativeness of a piece of research depend on such characteristics as the importance of the question that was asked and the care taken in the design and measures, rather than whether the answer to the question turned out to be yes or no. The rationale for this curious value system was stated by no less portentous an authority than the *APA Publication Manual*:

Negative results lacking a theoretical context are basically uninterpretable. Even when the theoretical basis for the prediction is clear and defensible, the burden of methodological precision falls heavily on the investigator who reports negative results. . . . Failure to replicate results of a previous investigator, using the same method but a different sample, is generally of questionable value. A single failure may merely testify to sampling error or to the conclusion that one of the two samples had unique characteristics responsible for the reported effect, or the lack of effect. (American Psychological Association [APA], 1974, p. 21)

Atkinson et al. make the obvious but important point that sampling variability is an equally valid explanation for the original, significant result. This passage was deleted from the third edition of the *Publication Manual*, but the field as a whole has not responded so nimbly.

One consequence of this practice is that published results will tend to represent a much larger proportion of Type I errors than is commonly supposed, especially considering the low power of most psychological experiments (Cohen, 1962). The problem is the more difficult to detect just because efforts at simple replication are unlikely to be published, whether successful or not. Greenwald (1975) believes there have been identifiable “epidemics” of Type I errors in the literature. One of the examples he discusses is the phenomenon of people exposed to information on a controversial topic more easily learning agreeable than disagreeable information. About 10 studies between 1939 and 1958 reported this effect, he says, and it made its way into the introductory and social psychology textbooks. Starting in 1963,

¹³Their unexpected statement that “The manuscript was more than three times as likely to receive a rejection recommendation when it reported statistically significant findings” (1982, p. 192) is actually the opposite of the pattern portrayed in their accompanying table.

however, there were consistent failures to replicate, despite attempts to explain the earlier findings in terms of uncontrolled factors.

This problem, to the extent that it exists, could evidently be addressed with a proposal like that of Walster and Cleary (1970) for an initial review of manuscripts with the results and discussion sections deleted.

9.5 Problems in Application of Statistical Inference to Psychological Research

The fundamental problems with the traditional theories of statistical inference in psychology derive from the frequentist interpretation of probability. The most direct of these consequences include the issues of behavioral versus epistemic orientation (decision versus evidential appraisal) and the orientation to the aggregate versus the individual. They are not necessarily problems for other fields, particularly applied research, but I shall focus here on the implications for basic psychological research.

9.5.1 Epistemic Versus Behavioral Orientation

Statements like Guilford's and Kerlinger's, quoted above, indicate very clearly their belief—and the sense of the statements is surely close to most psychologists' understanding—that we need statistical inference fundamentally for epistemic purposes, for the evaluation of hypotheses, and hence, if either the Fisherian or the Neyman–Pearson rationale for significance testing were relevant to psychology, it would clearly be the former. And, indeed, one searches the literature in vain for any argument in favor of the Neyman–Pearson approach, *as against* the Fisherian, in psychological research. If such arguments were to be found, they would presumably be given in the statistics textbooks; but there, however careful the exposition of the Neyman–Pearson doctrine, the rationale presented seems inevitably to be Fisherian.

Kempthorne (1972), in an essay in honor of George Snedecor (who was nearly 90 at the time), notes that the Neyman–Pearson theory is seldom practiced as it is preached, that research workers do not in fact take the decision orientation seriously.

From the viewpoint of the Neyman–Pearson theory of testing hypotheses—or as this author prefers, the Neyman–Pearson theory of accept-reject rules—an inspector is not permitted the following thought process. Suppose two particular data points D_1 and D_2 fall in the rejection region of size $\alpha = 0.05$. Suppose also that D_1 falls in the region of size $\alpha = 0.01$ and D_2 does not. Then it is very natural to take the view that D_1 disagrees with the null hypothesis more than D_2 . But to use phraseology that is becoming current, this would be an *evidential* conclusion. It appears that no such conclusions are permitted in the Neyman–Pearson *theory*. Indeed, it can happen that a sample point is in the rejection region of size 0.01 and is not in the rejection region of size 0.05. It may be true that those who use the Neyman–Pearson theory will reach the evidential conclusion above, and indeed many of the ideas of the theory have been taken over and used in an evidential way. But nothing in the Neyman–Pearson theory permits this activity. (pp. 179–180)

I hazard the opinion that Snedecor's *Statistical Methods* has had some appeal to scientists and has not been modified in basic outlook by the development of decision theory, because decision theory deals with problems that are so simple (e.g., how to approach the problem of making scrambled eggs) and so simplified as to have no essential relevance to the problems of research and development. (p. 182)

To support the evidential interpretation of Snedecor and Fisher, Kempthorne constructs a logical measure D of the distance of the data from the model, as Spielman (1974) does also, and then argues that since the p integral is a monotonic function of D , it may reasonably be taken as an ordinal measure of it, and hence as an index of support. That may well, in fact, be the best defense that can be offered for the contemporary practice of significance testing, but it is not a strong one. Oakes (1986) points out that the relationship between effect sizes and their associated probabilities is often such that trivial increases in D correspond to what are regarded as huge increases in significance.

The evidential interpretation of significance levels offers an irresistible target for doctrinaire frequentists. In the first place, if we are prepared to interpret probabilities evidentially, there is no apparent obstacle to Bayesian statistics, and we might as well avail ourselves of the substantial advantages of that approach to be able to evaluate hypotheses directly, rather than constraining ourselves artificially to probability statements about the data. It was to avoid such an interpretation of probability, after all, that the decision theory was designed. Spielman (1974) adds the more subtle criticism that such use makes the rest of the evidential context relevant. The credibility of the hypothesis is not just a matter of statistical significance but also of the prior credibility (not necessarily just a subjective degree of belief) of the hypothesis relative to its alternatives. Spielman argues that there can be no statement of absolute support outside of this context.

But these arguments can also be challenged in turn. We may preserve the evidential interpretation of significance tests and still allow other pieces of evidence as relevant, too, provided simply that we refrain from setting up the test result as a summation of *all* the evidence, and treat it instead as merely one more isolated piece of evidence to be added informally to the rest. But that sort of provisional attitude does not match very closely the way significance tests are regarded in practice.

9.5.2 *The Literalness of Acceptance*

A further consequence of the behavioral orientation of the Neyman-Pearson theory is the ambiguity of the concept of acceptance or rejection when applied to a scientific or statistical hypothesis. Rozeboom (1960) complained years ago that a hypothesis is not the sort of thing that can be accepted or rejected like a piece of pie offered for dessert, "thanks" or "no thanks." And Gière (1976) reminds us that, though the notion of "accepting H " would apparently commit us to acting always as though H were true, in practice there will be risky contexts where we shall renège on that commitment, in violation of the rules. Even Neyman and Pearson themselves appear

to have had some trouble with the notion of literal or unvarying acceptance or rejection of hypotheses in accordance with the dictates of a test.

There are two possible alternative responses: We can either reject some of the time, but not all of the time, when the result is significant; or we can always reject, but, as it were, with mental reservations. Neyman and Pearson implicitly opted for the former approach. In one of their early papers, they remarked that if a result falls in the rejection region, “We shall probably decide to reject” (1933a, p. 303). They even went so far (1933b) as to give formal recognition to the problem in defining a “region of doubt” as a third alternative to acceptance or rejection; but this suggestion, though it crops up in occasional textbooks, has never become part of the official doctrine. Pearson (1962) later acknowledged that when we get down to any *real* decision, we shall need to take into account many more or less subjective, nonquantifiable factors in addition to the significance level. Likewise Kempthorne and Doerfler (1969) argue that:

In fact, it appears that the choice of decision should be based only slightly on significance levels, and largely on its relation to the size of the experiment, the magnitude of effects which would be interesting, and on one's personal probability of the existence of such magnitudes. (p. 247)

But there is surely no way for the Neyman-Pearson theory of hypothesis testing to accommodate such hedges. If we do not *always* reject when the result falls in the 5% region, and only then, the 5% level has no determinate meaning. If the hedge is meant to acknowledge that other factors will sometimes enter into our decision to accept or reject, it may be appropriate; but then we cannot claim to be using the 5% or any other specifiable level; there is no longer any way to calculate the long-run frequency of false rejections.

Gièvre, on the other hand, elects the second alternative. He proposes to give meaning to the acceptance or rejection of a hypothesis by making explicit the scientific context in which the act is performed. The context of scientific inquiry—unlike, say, the context of quality control acceptance sampling in manufacturing—is such as to make acceptance or rejection necessarily provisional, subject to reversal in light of future revelations. The definition of that context, as he recognizes, lies beyond statistical theory and formal analogies. But perhaps even in Neyman-Pearson theory, we could preserve the meaning of error rates if we were willing to allow this tentative cognitive act as an instance of inductive behavior.

9.5.3 *The Individual Versus the Aggregate*

It is commonly assumed that expectation in the long run justifies choices in individual cases. That is indeed the rationale for the Neyman-Pearson theory, which is focused exclusively on long-run average error probabilities. The immediate interest of psychological research workers, however, is very clearly on individual tests, and not on a hypothetical aggregate. What we want to know is the probability that the population mean lies in *this* interval, not in *such* an interval as this.

There are many examples to show, however, that the long-run expectation cannot be relied upon to justify decisions on an individual basis. One is the Petersburg problem; another is Peirce's example of the two decks of cards; Hacking (1965) offers others. E. T. Jaynes (1976) sums up the general situation for the scientist.

Our job is not to follow blindly a rule which would prove correct 90% of the time in the long run; there are an infinite number of radically different rules, all with this property. Our job is to draw the conclusions that are most likely to be right in the specific case at hand; indeed, the problems in which it is most important that we get this theory right are just the ones (such as arise in geophysics, econometrics, or antimissile defense) where we know from the start that the experiment can *never* be repeated. . . .

This does not mean that there are no connections at all between individual case and long-run performance; for if we have found the procedure which is "best" in each individual case, it is hard to see how it could fail to be "best" also in the long run.

The point is that the converse does not hold; having found a rule whose long-run performance is proved to be as good as can be obtained, it does not follow that this rule is necessarily the best in any particular individual case. One can trade off increased reliability for one class of samples against decreased reliability for another, in a way that has no effect on long-run performance; but has a very large effect on performance in the individual case. (pp. 200–201)

The fact that the significance level refers to an average over a long-run aggregate rather than to the individual test has been the source of some unease among psychologists and has led to a controversy over the proper unit for determination of a significance level. A typical problem is posed by a matrix of 5000 correlations (a 100-item test would yield 4950 different correlations). If they are individually tested, we can expect 250 to be significant even if all 5000 nulls are true. The situation is a good one for showing up our uneasiness with the long-run Type I error rate as our criterion. Accustomed as we are to inspecting hypotheses individually, we are unsettled by the prospect of making over 200 mistakes in a batch. The possibilities for error are less disturbing when presented serially.

Work on the problem of multiple comparisons goes back at least to Duncan's master's thesis in the late 1940s,¹⁴ though debate on the general problem was stimulated by Ryan (1959), who attempted to handle the problem by arguing for the experiment rather than the hypothesis as the appropriate unit for computing the significance level. If, as in Dunn's procedure, the significance level is set at .05 for the whole matrix of 5000 correlations, the criterion for individual tests will be very much more stringent. Wilson (1962) countered, however, that while controlling the error rate per experiment will indeed result in fewer errors per experiment in large experiments, it will lead to more errors per hypothesis in small experiments, as

¹⁴There would appear to be an interesting chapter to be written on the development of post hoc tests. I am thinking particularly of the intriguing circumstance that, with the notable exception of the Scheffé test, these have some very obscure origins: the journal *Euphytica* (for the Newman-Keuls), the *Virginia Journal of Science* (for the Duncan), and a never-published manuscript by Tukey.

compared with larger experiments. In Ryan's scheme, comparability with respect to hypotheses would be lost; we would have to know how many hypotheses were tested in a given experiment before we could compare results across experiments. There is, moreover, no compelling logical reason to draw the line at the experiment as the unit, rather than the experimenter—the whole corpus of experiments performed in an investigator's lifetime—or in the entire community of experimental psychologists. Wilson himself opted for a conceptually defined unit; in some cases a comparison of five groups might be thought of as a single hypothesis, whereas in others the 10 pairwise comparisons would be the most relevant hypotheses; this unit, known as the familywise error rate, has gained wide acceptance.

All of the solutions are actually unobjectionable on logical grounds, including the original error rate per comparison. Ryan's move appears to be primarily motivated by discomfort with its implications over the long run; but it might be said, on the contrary, to be a virtue of the per-comparison criterion that it prevents us from dodging those implications. An alpha of .05 means simply that we can expect 5% of our tests of true nulls to result in erroneous decisions, whether we do all the tests at once or across a lifetime. Strict adherence to a Dunn-Bonferroni or similar adjustment would have mainly the effect of discouraging large-scale studies and encouraging piecemeal publication. It is also relevant here to keep in mind the point made by Spielman (1973) and Pollard and Richardson (1987) that, regardless of what α or β levels we set, we can never know our actual proportion of errors unless we know the proportion of true hypotheses in the reference class of hypotheses we test—as of course we never do.

The practice of adjustment for multiple comparisons thrives today mainly among journal reviewers who know nothing of statistics, but who want to make it appear that they have reviewed the Methods section. It provides an easy boilerplate criticism of almost any article. Although many journal editors are by now alert to the foolishness of arbitrarily constraining power, they often appear reluctant to overrule their reviewers en masse, so the demand seems destined to continue.

9.5.4 *The Paradox of Precision*

The next two problems result from asymmetries between the null and alternative hypotheses.

The first is the problem of testing a point null against a diffuse alternative. So long as the null represents only a zero (or some other) point, any other point, no matter how close, counts as an alternative. Suppose we are trying to determine whether a given coin is biased; the probability with which it yields a head will be some value in the continuous range from 0 to 1, of which the hypothesized value, .5, represents but an infinitesimal point. If there is any bias, however slight, it will show up as significant in a sufficiently large number of trials. Thus we arrive at an insight into Bernoulli's Theorem which is opposite to the usual intuition:

If the chance of an event occurring in a single trial is estimated to be p' , then the probability that the ratio of the number of times the event occurs to the number of trials differs from p' by an amount of any significance, however great, can be made as near certainty as desired by increasing the number of trials. (Spencer Brown, 1957, p. 90; original in italics)

We might call this the perverse form of Bernoulli's Theorem.

Now, Bakan (1966) and others have pointed out that there are few, if any, exact true nulls in nature. Consequently, we can virtually be assured of significance, whatever hypothesis we test, with a large enough sample. The prospect of "too large" a sample has led some writers (e.g., Binder, 1963; Hays, 1963) to recommend that investigators specify a size of effect they were interested in detecting, prior to doing the experiment. If we are prepared to say with what probability we should like to pick up effects of a given size, then the requisite sample size can be calculated. Research psychologists have generally been reluctant to adopt this solution, however, because it entails a nakedly personal judgment about how big a difference must be to matter. There are not, and could not be, any uniform, objective criteria for what size effect is theoretically interesting or important. In fact, it is largely to avoid this particular judgment that psychologists have relied so universally on significance tests, to answer questions which arguably are about *substantive* significance. Research practice changed in the late twentieth century, when the federal government began requiring power analysis for grants. Psychologists were saved from exercising judgment by Cohen's (1962, 1977) classification of effect sizes, though he himself cautioned that an effect size that would be considered small in some contexts might be considered medium in others.

Perhaps the neatest way of demonstrating just how we rely on significance tests to provide substantive answers is the following hypothetical example: Suppose we hypothesize a difference between the means of two populations on some dimension. Men and women, schizophrenics and normals, experimental and control, however the populations we want to compare, may be conceived, imagine that this time we have, not sample means, but the figures for the entire populations. There can thus be no question at all of statistical significance, no question of whether the difference could have occurred "by chance." We have the ultimate answer to our question. Now, in populations with even as many as a thousand members, it is virtually certain that the means will not be identical, down to the last decimal point. But, given that there is a difference between means, how big does it have to be to matter, to make a difference to us? We have become so accustomed to treating statistical significance as real significance that we have almost succeeded in concealing from ourselves that the answer to any such question, whether "there is a difference," involves an essentially personal judgment. It is of course thoroughly paradoxical that we should feel there was something insufficient about an ultimate answer, with no possibility of a p value; but the degree to which we are disconcerted by the implications of complete knowledge of population values is a measure of the extent to which we rely—inappropriately—on statistical significance to tell us, indirectly, what size effect is important.

Nor is the situation necessarily hypothetical. Conrad (1979), in an excellent study, tracked down all deaf school leavers in Britain aged 15–16—there were fewer

than 500—classified them in various ways and proceeded to do chi-square and other tests of significance. But such tests were totally meaningless as indicators of an association in the population: He had the population data in front of him. This example also makes clear the fallacy of supposing the “true” population to be some hypothetical aggregate encompassing deaf school leavers in different countries for all time: Deaf education differs distinctly between England and the United States, and in both countries it is undergoing noticeable change from one decade to the next.

Significance levels, it is true, vary roughly with the “distance” of the data from the model. But the problem with them in the present context is that they are also very heavily influenced by sample size. So long as we are unable or unwilling to specify a point rather than a diffuse alternative—or, equivalently, to specify the size of effect we are interested in—our significance tests are groping enterprises, even among fishing expeditions. If we have no idea what we are trying to catch, we have no way of knowing what size hook to use. A very large sample will reject anything but a perfect null; such a sensitive testing procedure, overaccommodating in the manner of the Bayesian’s continual revision of probabilities, will always lead to fine adjustments in our parameter estimates. A very small sample, on the other hand, will fail to detect any but the hugest effects; in the limit of insensitivity, it will never yield to alternatives except by accidents of chance.

Now we are led straight into a paradox. In general, by controlling sample size, we can virtually assure either rejection or acceptance of the hypothesis tested. But situations where we have such thorough control over the outcome do not look at all objective or scientific. In order to make the process look as though we are discovering something independent of ourselves, rather than merely forcing the desired conclusion, we have to make it appear that we had nothing to do with the choice of sample size. More precisely, we have to insure that the sample size falls within some intermediate range, so as to foil confident prediction; then we leave the outcome more or less to “chance,” allowing the exact size to be determined by factors other than what we wish to demonstrate. If extremal sample sizes happened to be the most convenient to collect, our predicament would be more embarrassing; in fact, significance tests would never have evolved as instruments for decision in the first place. The paradox of sensitivity would disappear altogether if we were prepared to designate a magnitude of effect of theoretical interest; but if we shifted our focus to the size of effects, as many of these authors have recommended, we are involved in old-fashioned problems of estimation, and large samples, and significance tests lose their relevance anyway.

9.5.5 *Identification with the Null*

In the usual practice of Fisherian significance testing, the theoretical hypothesis, the one that is actually believed, is represented by the diffuse alternative, and the point null represents the denial of any such effect. Thus significant results count in favor of the hypothesis, and we have a rejection-support (r-s) scheme, to use Binder’s

(1963) terminology. Occasionally situations arise, however, where it is logical to identify the theory with the null; the theory itself makes the claim of no effect. A common example of such an acceptance-support (a-s) scheme is the matching of two groups on background variables; we want to be able to assert that they do not differ.

The problem is most acute in Fisherian theory, which does not sanction acceptance of the null hypothesis (“proving the nonexistence of a difference”). Neyman-Pearson theory, which does not single out one hypothesis as the null, is slightly better off in this respect. Neyman (1942) contends that it is in principle arbitrary which hypothesis is designated as the null; since, however, the theory still treats errors asymmetrically, he suggests identifying the hypothesis tested as the one for which false rejections would be more costly. In psychological research the concept of costs of false rejections is a little vague, but it has generally been agreed, in the interest of scientific conservatism, that it is worse to declare the existence of an effect when there is none than to fail to discover a true difference. This resolution, at any rate, enabled Neyman-Pearson theory to be mapped onto Fisherian practice. The freedom, within Neyman-Pearson theory, to test either hypothesis or its denial is very seldom exercised, for the reason that it pertains only to point alternatives. A sampling distribution cannot be derived from a diffuse lumpen-hypothesis.

The result has been that our only accredited means of denying a difference has been to assert that the difference failed to reach significance. The procedure is awkward and unsatisfying, first because it is only the negation of a negation (failure to find evidence against the assertion of no difference), and second because it appears rather strongly biased in favor of the tester, with 95% (or 99%) of the distribution of outcomes counting for the null. Logically, it would be possible to designate a narrow band (5%, 10%, 20%) around the null point and assert the null only if the outcome fell in this center portion of the distribution, but the practice has never caught on. It would be an effective lesson, however, regarding the variability of small samples.

9.5.6 Random Sampling from Hypothetical Infinite Populations

Statistical inference of any kind rests crucially on the assumption of a random process; without such a process, an inference may be probabilistic in the strictly epistemic sense, but it is not statistical. Thus, from a logical point of view, the first issue in understanding statistical inference is to establish which process is to be conceptualized as a random process, subject to statistical modeling. There are several possible answers. In the first place, the population distribution itself is conceived as the result of a random process—for instance, the Bernoulli process underlying the normal curve of errors of measurement. At one time, this process was the focus; the assumed probability distribution was used as the basis for rejecting discordant

observations. In modern significance testing, however, it has become incidental; it makes no difference how the population values might have been generated, so long as they conform to the law of distribution assumed for them. The random process actually modeled in significance testing is either sampling from a population or assignment of individuals to treatment conditions, more often by far the former. The probability statements issuing from the model are distribution statements about aggregate outcomes—for example, the statement that the long-run behavior of (some particular aspect of) random samples from a certain population is governed by an F distribution (though a finite number of samples may exhibit any distribution whatever over a nonnegative range).

The peculiarity of such a model for psychology, of course, is that we only observe a single individual (a single sample), and there is no real aggregate prediction. In other applications of probability, as in physics or epidemiology, we observe an aggregate of individuals, albeit finite, which gives us a (probabilistic) check on the adequacy of the model. In psychology, however, where all but one are imaginary, we have no very good indication how appropriate our assumed model may be. It is somewhat as if we hypothesized a particular distribution of raindrops over a certain area and then accepted or rejected the model on the basis of having observed the location of a single drop.

The situation is complicated still further by the fact that, under the sampling paradigm, the operation which was supposed to be the random event, namely, the drawing of a sample, is in practice ill defined. Random sampling is ordinarily defined as a scheme that gives every possible sample an equal chance of being selected, and it ordinarily involves the devising of an indexing system for elements of the population. “In practice,” as Gière (1976) notes, “such a strong conception of random sampling is unnecessary. It is usually sufficient that the sampling mechanism exhibit no bias for or against individuals with characteristics relevant to the hypothesis in question” (p. 90). There is scarcely any pretense, however, to random sampling in psychology, even in this liberalized sense. As Nunnally (1960) says in a discussion of what he and others have called “the sampling fallacy,”

We should not take the sampling notion too seriously, because in many studies no sampling is done. In many studies we are content to use any humans available. College freshmen are preferred, but in a pinch we will use our wives, secretaries, janitors, and anyone else who will participate. (p. 645)

Guilford (1942) dignify this process with the name of “incidental sampling.” It is not at all clear, however, that such “incidental sampling” will do as a stand-in for true random sampling.

Are some or all of the specific benefits of probability sampling available regardless of whether the sample is a probability sample? We doubt it. Statistical inference depends on a statistical theory, but to be applicable the theory also depends on certain empirical operations in research. To ask whether a given result could be generated by a random process model in the absence of a random process in the generation of the data is simply to raise an irrelevant question; an absolutely crucial feature of the application of the model is missing. (Morrison & Henkel, 1969, pp. 134–135)

Although populations, and therefore samples, were once real (and in survey research still are), the Fisherian revolution made both the population and the random sampling hypothetical at the same stroke. Recall Fisher's formulation of the hypothetical infinite population as a fictional superstructure invented for the purpose of "reducing" the sample data. A given sample, however it is obtained, is treated *as if* it had been randomly drawn from some population, which is then defined by reference to the sample.

However literally Fisher intended the hypothetical character of the population to be taken, research workers have generally tried to interpret it in terms of real aggregates, and not a few statisticians support the practice. Thus Cochran and Cox (1950) say:

The hard fact is that any statistical inference made from an analysis of the data will apply only to the population (if one exists) of which the experiments are a random sample. If this population is vague and unreal, the analysis is likely to be a waste of time, at least from the strictly practical point of view. (p. 411)

In her classic text *Statistics for Sociologists*, Margaret Hagood (1941) struggled with the meaning and relevance of the hypothetical infinite population.

The fact of change in social and cultural phenomena renders unrealistic any conception of identical repetition of the complex of factors conditioning characteristics such as fertility and level of living. The concept of the universe of possibilities—that is all possible sets of measures on fertility and level of living that could possibly be produced in the thousand rural counties of the United States under conditions exactly similar to those of 1930—the concept has neither a realistic counterpart nor a readily imaginable counterpart. To what, then, does the variation expected from random sampling from such a universe of possibilities correspond? Only a feat of imagination involving an infinite prolongation of a present moment, where conditioning factors remain the same but "chance" factors continue to produce random variation can supply the answer. With this done, the observer sociologist along with the experimentalist still faces the problem of interpretation of the chance variation—with the alternatives of ascribing it to the present limitations in knowledge or to the statistical nature of the occurrence of events. (p. 430)

At present the sociologist must face the fact that the postulated, hypothetical, infinite universe of possibilities, concerning which he tests hypotheses to establish the "significance" of his results, is merely a logical structure, for which he can offer no real counterpart in his research situation. Then what is the utility of such a construct and of the tests of significance based upon it? The answer to this question is not perfectly clear at the present stage of the application of statistical methods to sociological research. (pp. 431–432)

9.5.7 Assumption Violation

The last problem to be considered here is a matter of practical concern, without particular implications for theory, but nonetheless important. It is simply the fact that in many analyses our models have become so elaborate and sophisticated that they cannot, by a long shot, be meaningfully tested. There are exceptions, of course, like the sign test and randomization tests in general; but the problem becomes acute

in multivariate models, which are enjoying wider and wider use. In the multivariate analysis of variance, for instance, the usual model assumes multivariate normal distributions with identical variance-covariance matrices and possibly unequal mean vectors. If covariates are included, then all the within-group regression coefficients must be assumed equal across groups, for each variate and covariate. If the design is factorial with n factors, the $2^n - 1$ terms must be assumed to combine additively; and so on.

For simple tests like the Student t , some well-known studies (e.g., Boneau, 1960) have shown that violation of some assumptions makes little difference in p values; but even here the more extensive studies (e.g., Bradley, 1959, 1963, 1964) tend to strike a more cautious note. Bradley (1964) argues that the limited, particulate studies focusing on one or two factors in isolation are misleading because the dozen or so factors influencing robustness interact with such complexity that virtually no general statements can be made about their effects even singly: the significance level, the location of the rejection region (one- or two-tailed), the number of samples, their absolute and relative sizes, the relative shapes and variances of the populations sampled, and interactions between the factors named, between violations, and between factors and violations. With respect to the t test in particular, Bradley's own investigation showed that:

Even under a liberal definition of robustness the two-sample t test is simply not very robust, or to put it more accurately, the test is drastically nonrobust under many of the conditions investigated in this study and relatively robust under few conditions. . . . When population variances are heterogeneous and samples are unequal in size [a ratio of 2/1 or 3/1], the distribution of the two-sample t (with perhaps certain rare exceptions of academic interest only) does not approach the normal-theory t distribution as its limiting distribution at $N = \infty$. Thus even a very liberal criterion of robustness may *never* be met at *any* sample sizes if population variances and sample sizes are sufficiently unequal. (A ratio of 2/1 for both sample sizes and population standard deviations was sufficient to produce drastic departures of ρ [empirical α] from α even at $N = 1024$ in the present study.) (1964, p. 109)

In the more elaborate multivariate models, the task of investigating the consequences of assumption violation is all but unimaginable. And until it is accomplished, we shall have very little idea what we are doing with such models. It is true that we need not be limited to these particular models; in some cases mathematical statisticians have developed more flexible tests—for example, allowing variation among within-group regression lines, or certain forms of nonidentity in covariance matrices—but these will remain unattractive to psychologists for several reasons. In the first place they require recourse to difficult mathematical expositions. In the second place, the specification of alternative, more realistic conditions, in either their form or degree, requires exactly the same kinds of choices as the specification of a point alternative hypothesis. It is easier to exercise the default option in the computer programs. Finally, perhaps, the search for more realistic models can be threatening because it brings to light all the inadequacies and arbitrariness in the old models.

In view of the discussion in the preceding section, it may be tempting to suggest that if the populations are only hypothetical anyway, then perhaps we might as well endow them with whatever characteristics we wish. In fact, the vague, hypothetical

nature of the relevant populations seems to have promoted just such an attitude. As Bradley (1968) emphasizes, in applied statistical contexts “assume” has practically come to be synonymous with “take for granted.” But the objection here is the same as that of Cochran and Cox to the hypothetical infinite population: If we want to get statements about the real world out of statistical analysis, we have to put some such statement in.

9.5.8 *The Impact of the Preceding Problems*

The severity of the problems attending statistical inference in psychological research raises the question of why we are apparently not suffering any consequences from them. The answer has to do with how far much of our research is removed from issues of real significance—in contrast with research on, say, the structural strength of an airplane wing. There are issues here of both substance and method.

Regarding the importance of the kinds of research questions psychologists tend to ask, Bakan (1967) likens our work to children playing Cowboys and Indians. Just as in their play children imitate all aspects except the essential work of cowboys—taking care of cows—so we teach our students to play scientist, to imitate the world of scientists in all but the essential respect, which is thinking and making new discoveries; we teach them the motions to go through, framing and testing hypotheses, and going through elaborate statistical analyses to demonstrate what is obviously true.

But even for research questions of importance, it is routinely very difficult to ensure the adequacy of both laboratory control and representation of real-world phenomena; as a consequence, the viability of alternative explanations is the rule rather than the exception. But the separation, the slack, between the world and the laboratory serves to cushion us from possible adverse consequences of misuse of statistics in our research.

Insofar as nothing of real significance continges on the results of our experiments, it would make very little difference what we did with our data, what agreed-upon rituals we performed, to reach our decisions about them. In this respect we are in a position somewhat like the man who snapped his fingers to keep the elephants away: We are shielded, by external circumstances beyond our awareness, from knowing whether our methods are any good or not. The elephants never approach to test the finger-snapping strategy, and reality never intrudes on our laboratory pronouncements.

9.6 Possible Responses By Frequentists and Bayesians

Over the years, a number of critics of significance tests (e.g., Binder, 1963; Oakes, 1986) have advocated the least radical alternative, which is the use of confidence intervals in their stead. Confidence intervals, it is true, obviate some of the problems

of the preceding section: problems of the literalness of acceptance of hypotheses, identification of theory with the null hypothesis, and the paradox of precision. Other issues—the epistemic versus behavioral orientation, the individual versus the aggregate, random sampling from hypothetical infinite populations, and assumption violation—remain as relevant for confidence intervals as for significance tests. Consequently, I believe Oakes exaggerates in describing confidence intervals as “infinitely preferable” (p. 66) to significance tests. Apart from pervasive problems of interpretation (Chap. 7), confidence intervals based on psychology size samples are typically wide enough to drive a truck through. That was the reason for the switch from estimation to significance testing in the first place. Large samples pose no paradoxes for confidence intervals as they do for significance tests, but psychologists are unlikely to start collecting samples of a size adequate to estimation. And, fundamentally, confidence intervals have no appeal for psychologists just because they don’t provide a yes-or-no answer.

The Bayesian approach can boast a slightly better scorecard, with claims of superiority on five of the seven problems discussed above. The other two—the problem of random sampling from a population (and associated issues) and the problem of assumption violation—are common to both the Fisher-Neyman-Pearson and the Bayesian approach.

As the frequency concept of probability is the major burden, in epistemic applications, for the traditional theories of statistical inference, so an epistemic concept is the major asset of the Bayesian theory. It is the frequency concept which necessitates restriction to forward probabilities, of which significance levels and confidence coefficients are examples, and most of the problems above are problems with these instruments.

The Bayesian approach retains the sampling theory core of frequentist inference—the likelihood—and sandwiches it between prior and posterior degrees of belief. We begin by sizing up all the relevant evidence and committing ourselves to some absolute degree of belief in the hypothesis in question and the various alternatives, provided that our allocation of belief is exhaustive. Thus in Bayesian statistics we can actually say, as Kerlinger (1973) wished to say, that there is such-and-such a probability, given our prior beliefs, that the hypothesis is true or untrue.

The statement, moreover, pertains to this particular hypothesis; it is not a statement about a long-run average over a sequence of hypothetical trials, or an aggregate of hypotheses. Ryan’s (1959) problem, of choosing the appropriate unit for a when multiple tests are made in the same experiment, does not arise in Bayesian statistics. Nor are we concerned with base rates of true nulls in the reference class of hypotheses we test, since we are not concerned with error probabilities in the first place.

As further consequences of using posterior probabilities instead of error probabilities as the outcome statement, the paradox of precision dissolves, as does the distinction between acceptance-support and rejection-support schemes. We can determine the posterior support for the null, as we saw in Chap. 8, by specifying a band around the point null value and reporting the posterior probability contained within that interval. Questions of power do not arise if we are not making decisions

about hypotheses. And large samples, instead of threatening us with the prospect of making trifling effects significant, merely add more weight, as indeed they should, to the sample results in relation to prior probabilities.

There are no puzzles about the meaning of acceptance or rejection of hypotheses in a scientific context, or questions about literal obedience to the dictates of a test. A Bayesian calculation yields simply a measure of support for the hypothesis in question.

Finally, to return to the wider epistemic implications of Bayesian statistics, the existence of information or impressions beyond what is given in the data poses no problem. If such information cannot be quantified in any other way, it may be measured by betting ratios and inserted into Bayes' formula. Indeed, in the view of some writers (e.g., Kyburg, 1976), this feature is decisive in favor of Bayesian statistics whenever we have information beyond what is represented in the sample. The Bayesians are at least alone in claiming to handle it, and other approaches can lead to absurd results if they ignore it (as in the confidence intervals for birth rates or product lifetimes). The frequentists may reply, of course, that the frequency data speak only for themselves and do not warrant any inference beyond themselves.

9.7 Toward Resolution: The Frequentists Versus the Bayesians

A final resolution will have to await some further developments covered in Chap. 10, but at this point, some preliminary observations can be made regarding the frequentists versus the Bayesians.

In the core of their calculations, the Bayesians and frequentists use the same classical probabilities. The issue is whether to convert these into degrees-of-belief probabilities expressing the degree of support for the hypothesis after the experiment, by supplying such degrees of support prior to the experiment, or whether to let these steps occur in an informal way. Clearly the *object* of Bayesian theory is geared to our needs in psychological research. Indeed many psychologists (like Kerlinger and Guilford) think a Bayesian-type theory is what they already have. The Bayesian theory, to the consternation of some of its advocates, supplies a rationale for the tacit inversion of significance levels, which is illegitimate within the frequentist theory. Specifically, when the prior distribution is diffuse, so that the sample represents the whole of the relevant information, then the posterior distribution of belief coincides mathematically with the sampling distribution, on which significance levels are based; hence the significance level could be interpreted, in a Bayesian way, as the probability that the population mean lay beyond a certain point, or the probability that the null hypothesis was true. Similar possibilities exist for nondiffuse priors. It is quite possible, in other words, even if it is totally against the rules of significance testing, informally and implicitly to invert significance levels, to get an epistemic probability statement. All that is required is informally and implicitly to supply a

prior distribution of belief. The question is what is gained, or lost, by doing this formally instead of informally.

In the history of American psychology so far, keeping as much as possible of the inferential process out of the formal theory has worked to the advantage of the frequentists. Essentially they have enjoyed the benefits of both Bayesian practice and frequentist theory. For, even though the Neyman-Pearson theory expressly forbids inversion of significance levels, in scientific applications, the temptation to inversion is so strong that the official interdiction has come to seem puritanical. Many psychologists, including some authors of statistics texts, seem not to have understood the dispute in the first place, but even those who have might be partly excused: For if legally the theory is still quite right in protesting its innocence of responsibility for such rampant abuse, we may still hold it partly to blame in its exquisite vulnerability to perversion.

The theory itself, as it is presently understood by psychologists, has the secure advantage of appearing by far the most objective of the available theories. It is presumed to be based on a scientific conception of probability, and its arbitrary conventions are by now well established. In addition, history has served to make the entire process of statistical inference or decision look more objective than it is. The effect here is wider than the history of statistics; it has to do with the kinds of distortion introduced into a theory by its epigons, who invariably make it more exaggerated and rigid. Keys (1972) puts the point more strongly, seeing the distortion as an actual inversion of the original doctrine.

A historical fact of some interest in this connection, which we cannot afford here to touch more than briefly, is that the founder of any religion is the man who tells how to undo the spells: but the church that establishes the religion, being so to speak its material embodiment, must, to maintain its worldly existence, present the founder's knowledge essentially in reverse, so that within its *corpus* the original knowledge becomes *secret*.

A relatively superficial and uncomplicated illustration of this is furnished through the teachings of the German philosopher Ludwig Wittgenstein, who may for the purpose of this example be considered as a minor christ. He taught that all philosophy, including his own, is nonsense, and that any order of existence other than the physical, although not unreal, is unspeakable.

For the purpose of this example we may take the philosophical school of linguistic analysis, or logical positivism, as Wittgenstein's established church. The teachings of this school at least suggest, if they do not actually say, that philosophy is the only way to talk sense, and that any order of existence, except the physical, is not unspeakable, but unreal. (1972, pp. 108–109)

Exactly the same thing has happened with Neyman-Pearson theory, which of course has very close ties with logical positivism. The founders taught that anything other than relative frequencies, although not unreal, is unspeakable, from the standpoint of the theory. But because the theory says, and can say, nothing about these subjective aspects—selection of hypotheses, their prior probabilities, costs of errors, magnitude of effects considered interesting, and so on—their followers have assumed, and have taught in turn, that they are unreal, in the sense that the theory is taken to

be exhaustive of the process of statistical inference or decision. And in practice our thought, true to form, has ceased to occupy itself with such concerns, as surely as if they had ceased to exist.

These banished aspects are now the source of part—if only part—of our troubles in statistics. Any theory which attempts to reintroduce them confronts the problem of appearing more subjective than the existing theory; and we cannot expect any change, at least in the moderate run, until such a challenger can come to be seen as less subjective than at present. Yet there are several possibilities. A simple example is the institution of conventional cut-offs for likelihood ratios (see Chap. 10). A more subtle possibility would be the recognition that in formalizing and making explicit a greater part of the whole process of inference, the Bayesians have a fair claim to greater objectivity. Thus, on one scenario, if the subjective aspects of statistical inference or decision were to be acknowledged, we could expect a quick switch to Bayesian methods: for the bitter pill of subjectivity can be swallowed only if heavily coated with convention.

The Bayesian claim of greater objectivity, despite its plausibility, can, however, be challenged in turn, as was noted in Chap. 8: for the Bayesian approach makes us equally vulnerable, by a different process, to the same fate of self-deception. Bringing subjective, hitherto unformalized aspects into the formal theory does indeed make them easier for us to see, but it makes it harder for us to see their arbitrariness. This is very much what happened in the advance of Neyman-Pearson theory over informal appraisal of data. With each step toward greater formalization, old, subjective choices are replaced by institutionalized convention; the obvious, personal subjectivity gives way to a collectivized subjectivity, which is then accepted as an ersatz objectivity; in the final step, we lose sight of the ersatz quality and suppose there to be no other form of objectivity. But there remains a residue; knowledge is not a fully formalizable process; there are, at least temporarily, points of diminishing returns, beyond which, the further we push the process, the less our formalization seems to capture.

It should also be noticed that we surrender our freedom in the process. As we write everything into the Bayesian (or some other) formula, we restrict our options for viewing our data; it all has to be done by theoretical prescription.¹⁵ All methodologies are institutions of control; that is their avowed purpose. The difference between the Neyman-Pearson and the Bayesian theories can be seen as that between the right and the left. The former, which established itself first, wants to keep everything clear, pure, and analytic, insisting on strict adherence to the rules and formulas, with no personal elements to enter in, to color or shape a conclusion. The latter, revolting against the puritanical tradition, aimed at a more human, personal philosophy, letting the People participate; but, though in theory probability is Your degree

¹⁵The results of the same process are already evident as a degenerative disease of the Neyman-Pearson theory: Too seldom any longer do we look at our data, or draw graphs; we are fast approaching a state where everything goes directly from an event recorder into the computer, and all we ever see, or report, is a p value. Thoughtful, innovative approaches to data analysis often mean a delay in publication, if not an outright rejection.

of belief, in practice Your beliefs must conform to the regulations of the theory—no vacuous or other nonadditive belief functions, and so on. To the old guard, of course, the Bayesian revolution looks like the destruction of the edifice of science itself, letting the whole class of rougher elements in by the front door, as it were—for “a government of men, not of laws,” in the catch-phrase of the right. But both sides undertake to prescribe for us the method of our inferences, always assuming some uniform, ideal conditions—for example, that all of our information bearing on a question consists of numbers—standardizing us, as knowers, in the name of science. No one, it seems, except Feyerabend (1975), has suggested that we may not need a government in the conduct of inquiry, any more than in the conduct of our lives.

It is debatable, in short, whether the Bayesian promise of yet more numbers and scientific rigor mortis represents an advance. The real advance may be to recognize that greater quantification is not necessarily an advance.

9.8 The Recent Integration of Bayesian Concepts and Methods Into Psychological and Medical Research

Starting around 1980, Bayesian concepts and methods made inroads into standard research practices which are striking for the degree to which they were unobtrusive and uncontested. These changes may look predictable enough to retrospect, but I, for one, did not see them coming. The driving considerations, as might be expected from the history traced in this book, were more pragmatic than philosophical.

9.8.1 *Hierarchical Models*

One prominent example is hierarchical regression models, also known as random coefficient models. These are two- (or three-, etc.) level models, where the coefficients at one level are random variables which vary as a function of still other variables. Thus, for example, in a longitudinal model, coefficients of a response over time (e.g., linear) can be modeled as varying as a function of some third variable, e.g., group membership. Or, putting time in the position of the third variable, we can see whether a correlation changes over time. Hierarchical models have become very popular in the last few decades.

Random effects models were known as far back as the 1930s, before any explicit introduction of Bayesian statistics. Hays (1963, 1973) was one of the few textbook authors to give them a serious presentation. It was a nontrivial topic, because the algebra was more complicated than for the more usual fixed effects. The practical rationale for their use was not compelling. Most of the categorical variables used as independent variables by psychologists are fixed, anyway (e.g., sex); even for the

textbook examples of random effects, like teachers, the assumption of random sampling was as strained as it was for the selection of participants.

The exposition of random effects models entailed an interesting delicacy: Mathematically there is no difference between a normal distribution of a random effect and a normal distribution of belief around a specified value. So frequentists discussing random effects often insisted that there was nothing Bayesian about their work. It was natural, in any case, that most of the distribution work on random effects was done by Bayesians. Such work (e.g., Lindley & Smith, 1972), however, remained theoretical until a way was found for estimation of the parameters. When that problem was solved (e.g., Dempster, Rubin, & Tsutakawa, 1981), and statistical programs were written to incorporate the new methods, hierarchical models won quick acceptance. And the distinctive Bayesian terminology (e.g., “shrinkage”) caused no particular concern.

In the incorporation into the mainstream of another Bayesian technique—multiple imputation of missing data—the fact that the mathematics was less well understood has had serious implications for practice, though they have not generally been recognized.

9.8.2 *Multiple Imputation of Missing Data*

The multiple imputation of missing data has always been treated as a Bayesian procedure, though there is no reason on the face of it why it should be intrinsically Bayesian. Schafer (1997) gives the cryptic explanation that the frequentist approach requires specification of the missingness mechanism and the Bayesian approach does not. But this is illogical: It is surely necessary for Bayesian as well as frequentist statistics. To see the source of this claim, we have to exhibit the mathematics.

If we assume that θ is the parameter of interest, that the variable y is distributed according to f , and that m is a missing data indicator distributed according to g , then the likelihood of the data is

$$L(\theta) = g(m|\varphi) f(y|\theta).$$

In Bayesian statistics if the prior probability of θ is $f'(\theta)$, the posterior probability is

$$f''(\theta|y) = \frac{f'(\theta)g(m|\varphi)f(y|\theta)}{\int f'(\theta)g(m|\varphi)f(y|\theta)d\theta}.$$

If φ is independent of θ , in other words if the missingness parameter is independent of the parameter being estimated—a condition that Rubin (1976) has defined as *missingness at random*—then the missingness mechanism g factors out of the integral, and cancels out of the ratio. Thus under a Bayesian solution the missingness mechanism is irrelevant, so long as the data are missing at random. This happy circumstance is an artifact of Bayesian posterior probabilities being conditional

probabilities and therefore ratios. As in this case a constant multiplier, of numerator and denominator, the missing data mechanism does not affect the ratio, as it does the absolute probability of the frequentist inference. There is nothing wrong with this result, so long as the conditional nature of the posterior probability is kept in mind; in practice, of course, it is not and is treated like any other absolute frequentist probability.

Another issue in the imputation of missing data concerns which variables to include in the imputation model. Specifically, the question is whether the variable being tested, say arm in a randomized controlled trial, should be included. Logically, it should not: In inserting a parameter, say Δ , for a treatment difference, we are estimating a nonzero value for the treatment difference, thus building into our imputation equation an estimated effect of treatment obtained in our sample. If we impute based on the null model, the resulting sample will be “biased” toward the null with respect to the original sample; but if we impute based on the nonnull model, our results will be biased away from the null. If we then find a significant effect, it will be partly because we put it there.

Interestingly, the authorities are unanimous that the parameter being tested must be included in the imputation model, so it is important to understand how this blunder could have come about. Rubin (1976) introduced multiple imputation in the context of estimating census data. For purposes of estimation, any information that will yield more accurate estimates is desirable and should be included. And there is no question of bias, because there is no hypothesis being tested; we are working with population data. Rubin would not have been inclined to consider what modifications may have been necessary for hypothesis testing, since hypothesis testing is not a natural practice for a Bayesian, anyway. When frequentists took over the practice, however, their focus was typically on hypothesis testing rather than estimation, and they failed to notice the circularity.

9.9 Postscript on Statistics in Medicine

The rather different history of statistics in medicine from that in psychology can be largely attributed to medicine’s having lagged psychology by about 50 years in becoming a statistical discipline. And this difference is related, in turn to the status of the respective disciplines. From its beginnings around 1900, psychology was desperate to place itself on a numerical footing for scientific respectability. The prestige of medicine was secure, although that prestige didn’t rest on the status of medicine as a science. Some, most notably Austin Bradford Hill, took notice of developments in statistics by Fisher and others, and urged medical research to haul itself into the twentieth century by embracing statistical method. But the resistance he encountered was fierce. Statistical method entailed essentially regarding all doctors, and all patients, as interchangeable; and many doctors clung to an image of themselves as practitioners of a skill that depended on detailed knowledge of

individuals. Hill wrote a series of articles for the *Lancet* on statistical methods, and they were published as a textbook in 1937, *Principles of Medical Statistics*. The first randomized controlled trial of a drug was not done until 1948 (Medical Research Council, 1948), funded by the Medical Research Council; that was still remarkably soon after the war. In the United States, when the FDA mandated randomized controlled trials in 1970 for approval of new drugs, the fantastic costs of the studies could be passed along to the consumer. That meant the choice of which drug to investigate was no longer a political one—it was up to the drug companies rather than, say, the Medical Research Council; it also meant, among other things, that drugs were favored which were for managing chronic conditions, and reinforced the trend to what Greene (2007) calls “prescription by numbers.” The fact that the cost of drug trials was now essentially unlimited meant that studies could be planned with binary outcomes. Such outcomes might have been natural in psychological treatment studies, but they had been avoided, both because they were expensive in terms of power and because methods like logistic regression could not be done without computers. They were thus not taught to psychologists much before 1970. But that is just when medical statistics took off, and such techniques have always been part of the statistics curriculum in medical schools. Similarly, a number of controversies that had played themselves out in psychology were not so charged by the time they entered medical statistics; the acceptance of Bayesian statistics was matter of fact by comparison.

References

- Adams, E. W. (1966). On the nature and purpose of measurement. *Synthèse*, 16, 125–169.
- Adams, J. K. (1955). *Basic statistical concepts*. New York: McGraw-Hill.
- American Psychological Association. (1974). *Publication manual* (2nd ed.). Washington, DC: Author.
- Anderson, R. L., & Bancroft, T. A. (1952). *Statistical theory in research*. New York, NY: McGraw-Hill.
- Anderson, T. W. (1957). *An introduction to multivariate statistical analysis*. New York, NY: Wiley.
- Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29, 189–194.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437.
- Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco, CA: Jossey-Bass.
- Baker, B. O., Hardyck, C. D., & Petronovich, L. F. (1966). Weak measurement vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 26, 291–309.
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425.
- Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, 101, 716–719.

- Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 70, 107–115.
- Blommers, P. J., & Forsyth, R. A. (1977). *Elementary statistical methods in psychology and education* (2nd ed.). Boston, MA: Houghton Mifflin.
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16, 21–33.
- Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, 57, 49–64.
- Boring, E. G. (1920). The logic of the normal law of error in mental measurement. *American Journal of Psychology*, 31, 1–33.
- Boring, E. G. (1966). Editor's introduction. In G. Fechner (Ed.), *Elements of psychophysics* (Vol. 1) (H. E. Adler, Trans.; D. H. Howe & E. G. Boring, Eds.), pp. ix–xvii). New York, NY: Holt, Rinehart and Winston.
- Bradley, J. V. (1959). *Studies in research methodology. II. Consequences of violating parametric assumptions—Fact and fallacy (TR 58-574(II))*. Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratories.
- Bradley, J. V. (1963). *Studies in research methodology. IV. A sampling study of the Central Limit Theorem and the robustness of one-sample parametric tests (AMRL-TDR-63-29)*. Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratories.
- Bradley, J. V. (1964). *Studies in research methodology. VI. The central limit effect for a variety of populations and the robustness of z, t, and F (AMRL-TR-64-123)*. Wright-Patterson Air Force Base, OH: Aerospace Medical Research Laboratories.
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.
- Brookhouser, P. E. (1984). Perspectives on adolescent concerns and needs. In G. B. Anderson & D. Watson (Eds.), *The habilitation and rehabilitation of deaf adolescents* (pp. 10–27). Little Rock, AR: University of Arkansas, Rehabilitation Research and Training Center on Deafness.
- Callahan, R. E. (1962). *Education and the cult of efficiency*. Chicago, IL: University of Chicago Press.
- Canard, N. F. (An X, 1801). *Principes d'économie politique*. Paris, France: Buisson.
- Carson, J. (1993). Army Alpha, Army brass, and the search for Army intelligence. *Isis*, 84, 278–309.
- Chambers, G. G. (1925). *An introduction to statistical analysis*. New York, NY: Crofts.
- Churchman, C. W. (1951). *Statistical manual: Methods of making experimental inferences* (2nd rev. ed.). Philadelphia, PA: Pitman-Dunn Laboratory.
- Clark, C. E. (1953). *An introduction to statistics*. New York, NY: Wiley.
- Cochran, W. G., & Cox, G. M. (1950). *Experimental designs*. New York, NY: Wiley.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.
- Conrad, R. (1979). *The deaf schoolchild*. London, UK: Harper & Row.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Cooke, D. H. (1936). *Minimum essentials of statistics*. New York, NY: Wiley.
- Croxton, F. E., & Cowden, D. J. (1939). *Applied general statistics*. New York, NY: Prentice-Hall.
- Danziger, K. (1984, June). *Educational administration and a critical shift in psychological research practice*. Paper presented at the meeting of the Cheiron Society, Vassar College, Poughkeepsie, NY.
- Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution* (Vol. 1. *Ideas in the sciences*, pp. 35–47). Cambridge, MA: MIT Press.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

- Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76, 341–353.
- Edwards, A. L. (1950). *Experimental design in psychological research*. New York, NY: Rinehart.
- Fechner, G. (1966). *Elements of psychophysics* (Vol. 1) (H. E. Adler, Trans.; D. H. Howes & E. G. Boring, Eds.). New York, NY: Holt, Rinehart and Winston. (Original work published 1860)
- Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London, UK: New Left Books.
- Fisher, I. (1913). *Elementary principles of economics*. New York, NY: Macmillan.
- Forward, J., Canter, R., & Kirsch, N. (1976). Role-enactment and deception methodologies. *American Psychologist*, 31, 595–604.
- Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, 103, 93–948.
- Gièvre, R. N. (1976). Empirical probability, objective statistical methods, and scientific inquiry. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 1. *Foundations and philosophy of statistical inference*, pp. 63–101). Dordrecht, The Netherlands: Reidel.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.
- Greene, J. A. (2007). *Prescribing by numbers: Drugs and the definition of disease*. Baltimore, MD: Johns Hopkins University Press.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1–20.
- Guilford, J. P. (1942). *Fundamental statistics in psychology and education*. New York, NY: McGraw-Hill.
- Guilford, J. P. (1956). *Fundamental statistics in psychology and education* (3rd ed.). New York: McGraw-Hill.
- Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hagood, M. J. (1941). *Statistics for sociologists*. New York, NY: Reynal and Hitchcock.
- Halphen, É. (1955). La notion de vraisemblance [The notion of probability]. *Publications de l'Institut de Statistique de l'Université de Paris*, 4, 41–92.
- Hansel, C. E. M. (1966). *ESP: A scientific evaluation*. New York, NY: Scribners.
- Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart, Winston. (2nd ed.), *Statistics for the social sciences*, 1973; 3rd ed., *Statistics*, 1981; 4th ed., 1988
- Hays, W. L. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart, Winston.
- Hornstein, G. A. (1988). Quantifying psychological phenomena: Debates, dilemmas, and implications. In J. G. Morawski (Ed.), *The rise of experimentation in American psychology* (pp. 1–34). New Haven, CT: Yale University Press.
- Jaggar, A. M., & Bordo, S. R. (1989). Introduction. In A. M. Jaggar & S. R. Bordo (Eds.), *Gender/Body/Knowledge: Feminist reconstructions of being and knowing* (pp. 1–10). New Brunswick, NJ: Rutgers University Press.
- Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2. *Foundations and philosophy of statistical inference*, pp. 175–258). Dordrecht, The Netherlands: Reidel.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press. (1st ed., 1939).
- Johnson, P. O. (1949). *Statistical methods in research*. New York, NY: Prentice-Hall.
- Kaye. (1998). *Economy and nature in the fourteenth century: Money, market exchange, and the emergence of scientific thought*. Cambridge, UK: Cambridge University Press.
- Kelley, T. L. (1923). *Statistical method*. New York, NY: Macmillan.
- Kempthorne, O. (1952). *The design and analysis of experiments*. New York, NY: Wiley.
- Kempthorne, O. (1972). Theories of inference and data analysis. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor* (pp. 167–191). Ames, IA: Iowa State University Press.

- Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika*, 56, 231–248.
- Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York, NY: Holt, Rinehart, Winston.
- Kevles, D. (1985). *In the name of eugenics*. New York, NY: Knopf.
- Keynes, J. M. (1936). *The general theory of employment, interest, and money*. New York, NY: Harcourt, Brace.
- Keynes, J. M. (1939). Official papers. Review of *A method and its application to investment activity* by J. Tinbergen. *The Economic Journal*, 49, 558–577.
- Keys, J. (1972). *Only two can play this game*. New York, NY: Julian Press.
- Koch, S. (1976). Language communities, search cells, and the psychological studies. *Nebraska Symposium on Motivation*, 23, 477–559.
- Kyburg, H. E., Jr. (1976). Statistical knowledge and statistical inference. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 2. *Foundations and philosophy of statistical inference*, pp. 315–352). Dordrecht, The Netherlands: Reidel.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22, No. 140, 1–55.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1–41.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Lonergan, B. J. F. (1970). *Insight* (3rd ed.). New York, NY: Philosophical Library.
- Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934–1945. *British Journal of Mathematical and Statistical Psychology*, 32, 151–178.
- Lovie, A. D. (1981). On the early history of ANOVA in the analysis of repeated measure designs in psychology. *British Journal of Mathematical and Statistical Psychology*, 34, 1–15.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151–159.
- McNemar, Q. (1949). *Psychological statistics*. New York, NY: Wiley.
- Medical Research Council. (1948). Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, 2, 769–782.
- Mills, F. C. (1955). *Statistical methods* (3rd ed.). New York, NY: Holt, Rinehart, Winston. (1st ed., 1924).
- Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *American Sociologist*, 4, 131–140.
- Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 105, 292–327.
- Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231, 289–337.
- Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society*, 24, 492–510.
- Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York, NY: McGraw-Hill.
- Nunnally, J. C. (1960). The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641–650.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.
- Partridge, E. (1966). *Origins* (4th ed.). New York, NY: Macmillan.
- Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics*, 33, 394–403.
- Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London, UK: Nelson.

- Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102*, 159–163.
- Polya, G. (1954). *Mathematics and plausible reasoning* (Vol. 1. *Patterns of plausible inference*). Princeton, NJ: Princeton University Press.
- Rice, S. A. (Ed.). (1930). *Statistics in social studies*. Philadelphia, PA: University of Pennsylvania Press.
- Rombauer, I. (1931). *The joy of cooking*. Indianapolis, IN: Bobbs-Merrill.
- Rothbard, M. N. (2006a). *An Austrian perspective on the history of economic thought*. (Vol. 1: *Economic thought before Adam Smith*). Auburn, AL: Ludwig von Mises Institute. (Original work published 1995)
- Rothbard, M. N. (2006b). *An Austrian perspective on the history of economic thought*. (Vol. 2: *Classical economics*). Auburn, AL: Ludwig von Mises Institute. (Original work published 1995)
- Rothbard, M. N. (2009). *Man, economy, and state: A treatise on economic principles, with Power and market: Government and the economy* (2nd ed.). Auburn, AL: Ludwig von Mises Institute. (Original work published 1962).
- Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57*, 416–428.
- Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthèse, 16*, 170–233.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.
- Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the “second discipline” of scientific psychology: A historical account. *Psychological Bulletin, 87*, 166–184.
- Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin, 56*, 26–47.
- Samelson, F. (1984, June). *On the impact of mental tests*. Paper presented at the meeting of the Cheiron Society, Poughkeepsie, NY: Vassar College.
- Sapolsky, A. (1964). An effort at studying Rorschach content symbolism. *Journal of Consulting Psychology, 28*, 469–472.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall.
- Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin, 72*, 214–228.
- Siegel, S. (1956). *Nonparametric methods for the behavioral sciences*. New York, NY: McGraw-Hill.
- Smith, J. G. (1934). *Elementary statistics*. New York, NY: Holt.
- Snedecor, G. W. (1946). *Statistical methods* (2nd ed.). Ames, IA: Iowa State College Press. (1st ed., 1937).
- Sorensen, H. (1936). *Statistics for students of psychology and education*. New York, NY: McGraw-Hill.
- Spencer Brown, G. (1957). *Probability and scientific inference*. London, UK: Longmans, Green.
- Spielman, S. (1973). A refutation of the Neyman-Pearson theory of testing. *British Journal for the Philosophy of Science, 24*, 201–222.
- Spielman, S. (1974). The logic of tests of significance. *Philosophy of Science, 41*, 211–226.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. *Journal of the American Statistical Association, 54*, 30–34.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York, NY: Wiley.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
- Student. (1908). The probable error of a mean. *Biometrika, 6*, 1–25.
- Theocharis, R. D. (1961). *Early developments in mathematical economics*. London, UK: Macmillan.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.
- Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. *Psychological Bulletin, 76*, 105–110.

- Velikovsky, I. (1950). *Worlds in collision*. Garden City, NY: Doubleday.
- Velikovsky, I. (1955). *Earth in upheaval*. Garden City, NY: Doubleday.
- von Mises, L. (1966). *Human action* (3rd ed.). Chicago, IL: Regnery.
- von Mises, L. (2007). *The historical setting of the Austrian school of economics*. Auburn, AL: Ludwig von Mises Institute. (Original work published 1969).
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100, 426–432.
- Walker, H. M., & Lev, J. (1953). *Statistical inference*. New York, NY: Holt.
- Walster, G. W., & Cleary, T. A. (1970, April). A proposal for a new editorial policy in the social sciences. *American Statistician*, 24(2), 16–19.
- Welkowitz, J., Ewen, R. B., & Cohen, J. (1976). *Introductory statistics for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.
- Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin*, 59, 296–300.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557–585.
- Yule, G. U. (1911). *An introduction to the theory of statistics*. London: Griffin.
- Yule, G. U., & Kendall, M. G. (1937). *An introduction to the theory of statistics* (11th ed.). London: Griffin.

Chapter 10

Recent Work in Probability and Inference



The present chapter has several purposes: to examine philosophical arguments on the nature of reasoning and inference, as well as selected psychological research data, by way of clarifying those concepts, and to survey selected other approaches to probability and inference, both by way of establishing what alternatives may be lurking outside the range known to psychologists and of throwing new light, in some cases, on existing theories.

10.1 Statistical and Nonstatistical Inference

10.1.1 *The Putative Philosophical Distinction*

The epistemological heritage of the seventeenth and eighteenth centuries found its ultimate expression in the logical positivism of the early twentieth century. Knowledge, according to the variants of this philosophy, consisted of propositions strictly bifurcated into analytic and synthetic. The former had shrunk to logic and mathematics, and the assimilation of logic to set theory undermined the distinction between these. As the source of the truth of such propositions and of our certainty of them, innate ideas had given way to social convention. The remainder of our knowledge claims (beyond “Here now blue,” etc.) represented empirical generalizations, which necessarily fell short of certainty.

The identification of uncertainty with probability and of probability with the mathematics of random processes carried the implication that all uncertain inference was fundamentally statistical—that all nondeductive inference could be modeled on or reduced to statistical inference. If the Rule of Succession had been discredited as resting on a metaphysical base, the formulation of new theories of statistical inference was being vigorously pursued. And if the mathematics were not always apparent, that was evidently just because in everyday contexts, we lack the

quantitative data and/or the inclination to do the actual calculations. But on this view, inductive reasoning in everyday contexts is just an informal, less precise version of what a statistician would be doing, and the latter in fact provides our model. This assumption is not always made as explicit as it is in Sarbin, Taft, and Bailey (1960), but it is implicit in the work of many writers (including, at least at times, Piaget) and might even be said to constitute the unreflective “folk wisdom” on the subject, at least among academic folk.

It is interesting to note that such an assimilation of all nondeductive reasoning to statistics rests implicitly on the extreme frequentist interpretation of probability, that of Reichenbach. Although Mises and other limited frequentists equivocate (recall Mises’ example of the tennis player), in setting aside everyday probabilities as nonscientific, they do not so easily permit the elision—as we saw, for example, in Cowles (1989)—from “All inductive inference is uncertain,” to “All inductive inference is probabilistic,” to “All inductive inference is statistical.”

The orthodox view of inference left logical room for its denial: that, with respect to either deductive or nondeductive reasoning, a purely formal analysis was insufficient in principle and that, with respect to so-called inductive reasoning, a statistical model was not generally appropriate. Although these alternatives were generally less well established and often remained only implicit, they were given formal recognition, in the philosophical literature, in terms such as Stephen Pepper’s (1942) distinction between multiplicative and structural corroboration. The former refers to the rather weak support available simply from compiling an aggregate of similar instances, whereas the latter proceeds from varied rather than similar observations, and its strength depends not on the number of cases studied, but rather on their ties to theory. L. Jonathan Cohen (1977) intends a similar distinction with his concepts of Pascalian and Baconian probability.

In the psychological literature, the distinction is represented, for example, in Yin’s (1989) philosophically less elaborated concept of statistical versus analytical generalization. The former, referring to statements about a population based on a random sample (i.e., statistical inference), corresponds in an obvious way with Pepper’s multiplicative corroboration and Cohen’s traditional Pascalian probability.

Possibly the most elaborate such distinction was that drawn by Bernard Lonergan (1957/1970) between classical and statistical heuristic structures. Lonergan, a Thomistic theologian, is rather far from positivism, and naturally would not be inclined to see all nondeductive inference as statistical. And, careful thinker that he is, his concept of statistical “heuristic structure” is not quite synonymous with statistical inference. Yet the close parallel he endeavors to construct between classical and statistical heuristic structures pulls him toward an endorsement of statistical inference as a complementary mode of reasoning.

Lonergan’s classical and statistical heuristic structures are grounded in systematic and nonsystematic process, respectively, with systematic process defined as follows:

1. The whole of a systematic process and its every event possess but a single intelligibility that corresponds to a single insight or single set of unified insights,

2. any situation can be deduced from any other without an explicit consideration of intervening situations, and
3. the empirical investigation of such processes is marked not only by a notable facility in ascertaining and checking abundant and significant data but also by a supreme moment when all data fall into a single perspective, sweeping deductions become possible, and subsequent exact predictions regularly are fulfilled. (Lonergan, 1970, p. 48)

Nonsystematic processes are created by violating the principles of systematic process. Whereas classical inquiry is concerned with *general* properties of a class, statistical inquiry concerns *coincidental aggregates*. A sequence of coin tosses is a clear example: It possesses a unity based on temporal succession, but there is no unity in terms of insight and intelligible relation. The concept of randomness is also intimately involved in nonsystematic process; Lonergan defines a “situation” as random (an epistemological, “achievement” definition similar to Fisher’s) “if it is ‘any whatever provided specified conditions of intelligibility are not fulfilled’” (p. 51). Thus, a nonsystematic process is characterized by randomness.

Systematic processes are best exemplified by those physical phenomena subject to modeling by differential equations, such as fluid motion or the movement of heavenly bodies. Concrete inferences from classical laws require (only) that we know

1. which laws are to be selected for the inference
2. how the selected laws are to be combined to represent the spatial and dynamic configuration of the concrete situation, and
3. what dimensions in the situation are to be measured to supply numerical values that particularize the selected and combined laws (p. 46).

Nonsystematic processes are represented by the weather, the stock market, birth and death rates, and the like. Concrete inferences here require a whole set of data for each new particular. It is in the nature of the difference between systematic and nonsystematic process, Lonergan suggests, “that astronomers can publish the exact times of the eclipses of past and future centuries but meteorologists need a constant supply of fresh and accurate information to tell us about tomorrow’s weather” (p. 51).

Nonsystematic process may be comprehended by its own special type of insight, however. Just as classical inquiry seeks insight into the *nature* of a phenomenon by abstracting from differences in measurements among particulars, demanding the qualification “other things being equal,” so statistical inquiry seeks intelligibility in the *state* of some phenomenon by prescinding from differences in particular frequencies of occurrence. The abstraction reached through statistical heuristic structure is the probability characterizing a given coincidental aggregate; it represents a number from which relative frequencies in a series cannot systematically diverge.

Classical procedures would yield *particular*, probably verified, conclusions about single events assigned a *unit* probability, where statistical procedures would yield *general*,

probably verified, conclusions about events as members of coincidental aggregates by assigning them *fractional* probabilities. (Lonergan, 1970, p. 68)

Constructing this comparison, however, makes Lonergan (evidently unwittingly) a Bayesian, in speaking of the probability of a conclusion in statistical procedures; by “probably verified,” he may intend either a Bayesian or, more likely, the everyday meaning. And, in fact, the probabilities reported by meteorologists typically represent not relative frequencies, but subjective assessments of evidence. There is no question, of course, that accumulating instances affords a sort of (weak) support for a proposition; but, whatever the utility and elegance of the distinction between systematic and nonsystematic process, it is not clear that the meteorologist or the stock market analyst employs a mode of reasoning fundamentally different from the astronomer’s.¹

In the psychological literature, the distinction between these two heuristic structures can be discerned, at least implicitly, in the work of Kahneman and Tversky (Kahneman, Slovic, & Tversky, 1982). Not surprisingly, these authors do not refer to Lonergan, and they do not use the term *heuristic of randomness*, but it is convenient to set up in parallel with their heuristic of representativeness. Under the latter, we treat an individual as representative of a larger class and make attributions on that basis. An epistemological gloss on Lonergan’s systematic process brings it closer: An intensionally defined class, exemplifying systematic process, is essentially constructed in such a way that an individual can be taken as representative of the whole. Whereas in coincidental aggregates, such as extensionally defined populations or strings of random events, we take individuals as they are, so to speak, in systematic process, we attach enough *ceteris paribus* and *mutatis mutandis* clauses to achieve by stipulation the necessary representativeness. By the heuristic of randomness, on the other hand, I shall mean essentially that we treat any given individual as a random sample from some suitably defined population and then apply the appropriate statistical formula (e.g., Bayes’ Theorem), exactly as if we were computers that had been programmed to calculate the required probabilities. The thrust of much of Kahneman and Tversky’s early research was to show that people use the heuristic of representativeness when, according to these authors, they should be using the heuristic of randomness.

¹I should acknowledge that the brilliant work of Stephen Wolfram (2002) has drastically transformed the study of nonsystematic process. Lonergan may be one of the few authors Wolfram doesn’t cite, but Wolfram’s analysis of nonsystematic process provides a model, in the form of cellular automata, and is not limited to modeling statistical frequencies. Wolfram does not claim, however, that his cellular automata support a model of *reasoning* about nonsystematic process.

10.1.2 Psychological Research on Reasoning in Statistical Contexts

It would be easy to misconstrue the purpose of this brief review. Although there is ample and diverse evidence by now that people do not adhere to statistical heuristics in problems where psychologists expect them to, or else they are not very good at doing the complex calculations in their heads, discrepancies between actual reasoning and a putative ideal cannot be used in any simple way to impugn the ideal. But neither, perhaps, can such discrepancies be used, with any greater logical warrant, to impugn human reasoning processes. Comparable discrepancies, some of them formally similar, are amply documented in deductive reasoning, but the indisputable correctness of the laws of logic does not necessarily make them an appropriate model or algorithm of thinking. If it did, we should expect Spencer Brown's (1972) elegantly simple "laws of form" to provide a still more plausible model for thinking than the ungainly taxonomy of syllogisms, but there is no reason to suppose that the truth-table logic of his algebra models reasoning, either.

This review will be part of a broader examination of the relation between reasoning and its formalizations, with the aim of throwing some further light on the concept of statistical inference.

10.1.3 Models of Inference

The assumption of "cognition as intuitive statistics," to use the title of Gigerenzer and Murray's (1987) book on the subject, can be traced rather directly to the pervasive externalization of knowledge in the eighteenth/century that was discussed in Chap. 2. The dissolution of the concept of causality, given formal expression in the concept of material implication, left the question of whether a bird has a head or a day has a sunrise on a footing with the question of whether a steamship has a flag. All characters were detachable from entities and from each other and were linked only through the superficial relation of statistical frequency. Any other basis for connection was either psychological or metaphysical, and by the early twentieth century, psychology itself would be pushed inexorably into metaphysics, by the positivist school of behaviorism.

So long as psychology shunned reference to consciousness as metaphysical and unobservable, any study of thinking and reasoning was naturally precluded. The emergence of cognitive psychology around 1970 was due, however, not to any humanistic overthrow of behaviorism, but to the ascendancy of the digital computer. Arising thus in the positivist mainstream, this discipline inevitably took deductive logic and probability theory as norms for human reasoning, and the salient problem became to account for departures therefrom.

In the analytic nature of logic and mathematics, only formal properties were presumed relevant to reasoning. The computer, limited in its operations to logical

manipulation of given quantities and labels, can take no account of content and context and thus models what was taken to be perfect inference. In the computer, Descartes' ideal of disembodied cognition achieves its ultimate embodiment, as it were: The computer is wholly material, but even so, it has no passions to distort its conclusions, its input devices yield a perfect copy, and it replaces thinking with calculation. Computers are accordingly interchangeable; there is no experiencing self with its unique perspective and history.

The discipline of cognitive psychology has thus been bound up from the start with the project of artificial intelligence, which, like statistical inference, was made possible by a general conception of knowledge as data. As Dreyfus (1979) notes, whatever plausibility that project can claim has its roots in the Greek separation of reason from the body and from the world of concretes. "In some other culture, the digital computer would most likely have seemed an unpromising model for the creation of artificial reason, but in our tradition, the computer seems to be the very paradigm of logical intelligence" (p. 231). Many of the problems encountered are accordingly the same. Of the four assumptions which Dreyfus lists as fundamental to the project of artificial intelligence, at least the last three are also relevant to the concept of statistical inference: (a) the biological assumption that the brain processes information in discrete bits, (b) the psychological assumption that the mind operates on discrete bits of information according to fixed rules, (c) the epistemological assumption that all knowledge can be formalized, and (d) the ontological assumption that reality is a set of isolable, independent facts.

The Laplacean urn model of induction was the first and prototypical example, but in time, the implications of the idea were developed explicitly that all mental operations could be modeled mathematically. Bayes' Theorem was taken as a model of all learning from experience, multiple regression was put forth as a model for clinical prediction, and factorial analysis of variance was suggested as a model of causal reasoning.

In its first phase, which may be coming to an end, cognitive psychology took as a principal object the discrepancies between actual human reasoning and these presumed logical and statistical norms. During this period, however, philosophy evolved also. Workers in artificial intelligence, in particular, were busy trying to simulate human cognitive performance. In the process, not only were the original models abandoned, but the distinction between ideal and actual, between philosophy and psychology, grew more vague. These latter developments will be briefly surveyed after a review of the traditional work and some of its implications for psychological research methodology.

Bayes' Theorem Bayesian statisticians have often claimed it as an advantage that their theory modeled human learning from experience in its progressive modification of personal probabilities by data, and they helped to initiate a new line of cognitive research in the 1960s. Neyman-Pearson theorists, disavowing inferential connections, were naturally less inclined to pursue the psychological isomorphism, and of course, the Laplacean Rule of Succession had been abandoned long before the emergence of the modern discipline of cognitive psychology. A Neyman-Pearson-type

decision theory has been taken as a model in signal detection theory (Gigerenzer & Murray, 1987), but Bayesians have held a near-monopoly in cognitive psychology.

As nearly unique and uncontested as it has been, we should not let the Bayesian claim to model the growth of knowledge pass unchallenged here. For the idea of a bit-by-bit accumulation of facts fits only a naive, mechanistic view of the growth of knowledge which not even a positivist would any longer endorse. As a representation of the history of science or of the ontogenesis of an individual's everyday understanding, the discontinuities under Neyman-Pearson theory actually come closer to the mark, except with the most brutally empirical "data" devoid of theoretical implications. With beliefs in which we have any investment at all, such as scientific or political theories, we tend to hold onto our current conceptions in the face of mounting evidence to the contrary, to a point where the anomaly becomes intolerable (a point that varies widely between individuals, according to their relative tolerance for anomaly versus ambiguity); then, there is a sudden shift to a whole new way of thinking.²

In the history of research on Bayesian information processing, Gigerenzer (Gigerenzer & Murray, 1987) discerns two distinct waves, differing in paradigm and conclusions. All of this work has been cogently criticized, going back at least as far as Acree (1978/1979); what I say here only brings together what has been developed at much greater length by Gigerenzer and Murray (1987), L. J. Cohen (1979, 1982), and Howard Gardner (1985).

The first wave comprised the "bookbag-and-pokerchip" experiments (cf. Edwards, 1968). Individuals were told, for example, that a sample had been randomly drawn from one of two bags containing red and blue chips, one bag contained 700 red and 300 blue chips, the other 300 red and 700 blue, the experimenter had selected a bag by tossing a fair coin, and sampling with replacement had drawn 8 red and 4 blue chips. The task was then to estimate the (posterior) probability that the bag sampled was the first. Most people responded in the range of .7–.8; few even approached the "true" probability of .97 yielded by Bayes' formula. The principal finding from this line of research was thus the phenomenon of "conservatism," by which it was meant that people tended to be unduly swayed by prior probabilities.

In contrast to these highly artificial problems, reduced almost to the numbers to be plugged into Bayes' formula, the second wave of research, done mostly by Tversky and Kahneman, presented story problems, and suddenly, conservatism was replaced by the opposite phenomenon of the "neglect of base rates." The difference in applying Bayes' Theorem to more real-life situations lies first of all in the assumption of random sampling. This assumption, which is required in all applications of statistical formulas as normative models for cognition, has been made routinely and cavalierly, if only implicitly, at least since the time of Price and Laplace. Its

²Chow (1988), in a curious inversion, argues that traditional significance testing provides the better model of learning because it represents the bit-by-bit accumulation of knowledge. (With similar logic, he argues for significance testing over estimation of effect sizes because the latter fails to provide the binary decision research workers are supposedly seeking.)

convenience, however, is not matched by its plausibility in everyday or scientific contexts: That is one fundamental problem with the concept of statistical inference.

A well-known experiment by Kahneman and Tversky (1973) is illustrative. People were presented with a thumbnail personality description, supposedly written by a high school psychologist on the basis of projective tests, which was more or less stereotypic of individuals in certain kinds of occupations: Tom W. was described as aloof and compulsive to a rather extreme degree and was said to be currently enrolled in graduate school. People were asked to rank nine fields of graduate specialization in the order of their likelihood as Tom's field of study. Another group, different but assumed similar, was asked to rate how similar Tom was to the typical graduate student in each of the nine fields, and a third group, not given information about Tom at all, was asked to estimate the percentage of graduate students actually enrolled in each of the fields. Now, what happened was that the likelihood of the various fields was judged essentially by the similarity between Tom and the typical student in the field: the correlation between likelihood and similarity rankings was .97. Kahneman and Tversky deprecate their participants' performance on the ground that they disregarded base rates and made their predictions strictly on the basis of representativeness, whereas a statistical approach, using the incidence of students in the various fields as prior probabilities, would have yielded substantially different results. Kahneman and Tversky, in fact, faulted their participants' performance on the additional ground that they accepted the data as given. They argue that the personality sketch should have been discounted in the first place, in view of the notorious unreliability of projective tests, and hence that predictions should have been even closer to base rates.

Statistical method, utilizing base-rate probabilities, would apply, however, only if the personality sketch at hand had been randomly sampled from the population of personality sketches of all graduate students. This was obviously not the case; the description was deliberately constructed to be typical of fields with comparatively few students; the experimenters did not even pretend to participants in their study that the sample was randomly chosen. In such a situation, the authors' reasoning was arguably inferior to their participants'.

The experiment was subsequently modified to take account of some of these considerations. This time, people were told that the sample was randomly selected. Base rates were made more salient by having the same people estimate both base rates and likelihoods, as well as similarities. And to highlight the feature of reliability, participants were told that students like themselves, on the basis of such thumbnail descriptions, made correct predictions in about 55% of the cases (another group was given the figure of 27%). The manipulation of expected accuracy had no effect on prediction; the orderings on the nine likelihoods were not significantly closer to the base-rate ordering under the low-accuracy conditions than under the high-accuracy condition. Participants did utilize base-rate data when no individuating information (personality sketch) was given; otherwise, they predicted from representativeness. Again, the key element for statistical prediction is the randomness of the sample, and it is surely questionable to what degree this piece of "information"

was registered by the participants or believed. It was, after all, false, and not very plausible at that.

Gigerenzer, Hell, and Blank (1988) manipulated the salience of the randomness assumption by having people in some conditions draw the description themselves from an urn. They found that they could produce base-rate (Bayesian) predictions under these conditions when people first “drew” (the “sampling” wasn’t random here, either!) uninformative descriptions, but that base rates were neglected, even with explicit “sampling,” if they first drew descriptions slanted to particular professions. In the latter case, as Kahneman and Tversky had found, people predicted from representativeness.

A second issue in the use of Bayes’ Theorem for reasoning in “real-life” applications is interpretation of the various terms of the formula, in other words, the problem of the ambiguity of the reference class. The difficulty is perhaps best attested by the fact that Kahneman and Tversky found they had to revise a simple problem several times to address this issue, just as they had in the Tom W. problem—in each case having imagined that their own initial interpretation was unequivocally correct.

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- (a) 85% of the cabs in the city are Green and 15% are Blue.
- (b) a witness identified the cab as Blue. (Kahneman et al., 1982, p. 156)

Information on the reliability of the witness was presented, and subjects were asked for the probability that the cab involved in the accident was Blue.

Kahneman and Tversky first supposed that it sufficed to say that the witness was 80% reliable; then, Bayes’ Theorem should yield for the probability that the cab was Blue (B), given that the witness said it was Blue (“ B'):

$$p(B|B') = \frac{p(B'|B)p(B)}{p(B'|B)p(B) + p(B'|G)p(G)} = \frac{8(.15)}{8(.15) + .2(.85)} \approx .41.$$

In other words, despite the reliability of the witness, the cab was still more likely to be Green than Blue. Respondents, “neglecting base rates,” gave a median and modal estimate of .80 for the probability that the cab was Blue.

The foregoing solution, as straightforward as it initially seemed, rests on some questionable assumptions: that Blue and Green cabs are equally likely to be involved in accidents, in hit-and-run accidents, and in hit-and-run accidents at night, and that the witness is equally reliable in identifying Blue and Green cabs. Variation of these conditions in the problem affected the answers given and, short of exhaustive specification of conditions (hit-and-run accidents at night, reliability of witness for each kind of cab), there is no uniquely defensible interpretation.

If there is an unambiguously correct solution to such textbook story problems, however—once the problem has been unambiguously formulated—there is no such clearly correct solution in real-life situations. In the former case, as Gigerenzer (Gigerenzer & Murray, 1987) and L. J. Cohen (1979) emphasize, the experimenters

have already stipulated, at least implicitly, what the relevant data are. In real life, on the other hand, just that determination would be the major part of the problem, assuming the relevance of Bayes' formula in the first place.

This is one of the reasons, in fact, why Cohen sees traditional, or "Pascalian," probability as irrelevant to most real-life situations. "Pascalian functions grade probabilification *on the assumption that* all relevant facts are specified in the evidence, while Baconian ones grade it *by the extent to which* all relevant facts are specified in the evidence" (L. J. Cohen, 1979, p. 389). His Baconian probability, as he calls it, has more to do with the weight of evidence, as Keynes and others used this concept, than with classical probability. "What determines the extent of the inductive support that identically favourable test-results give to a hypothesis," for example, "is the structure of the test carried out, not the number of times the same test-result has in fact been repeated" (p. 391). Such multiplicative, in contrast to structural, corroboration, to use Pepper's (1942) terms, often carries little weight. As Cohen points out, the sheer datum of the number of Blue and Green cabs in the city would easily be outweighed by the knowledge, for example, that the two companies tended to operate in different parts of the city, and the presence of such complicating information or impressions is the rule rather than the exception.

Whereas Cohen criticizes the Bayesian model of thinking for using the wrong concept of probability, Gigerenzer challenges it more nearly on its own turf. His criticism, faulting the Bayesian model for its inflexibility in tying decision criteria to base rates, is no less embarrassing for being less radical. Following an earlier analysis by Birnbaum, he construes the cab problem in the Neyman-Pearson framework, where the stimuli of Green and Blue cabs are assumed to produce overlapping normal distributions along a sensory continuum; above a certain threshold, a witness reports Blue, below it Green. If the witness is aware that only 15% of the cabs are Blue, a strategy of minimizing both types of error would lead to a very stringent " α ," or $p(\text{"B"}|G)$, reporting Blue only if the stimulus were very far out on the Blue end (and so much farther out still on the Green tail). Under these conditions, according to Gigerenzer, $p(B|\text{"B"})$ turns out to be .82, or just what Kahneman and Tversky's subjects said. If the witness is ignorant of the relative frequency of Green and Blue cabs, a strategy of answering Green and Blue equally often leads to $p(B|\text{"B"}) = .28$, a result which would have been interpreted as conservatism.

The phenomena of conservatism and neglect of base rates thus represent poles of a continuum determined by the relation of α to β . Gigerenzer argues that it is reasonable to let the decision criteria vary with the base rate, or prior probability, and that that is what people tend to do; but.

The so-called normative answer by Bayes' theorem implies that the ratio of the hit rate to false alarm rate is *independent* of the ratio of the base rates, as can be easily seen from the odds version of Bayes' theorem:

$$\frac{p(\text{B}|\text{"B"})}{p(\text{G}|\text{"B"})} = \frac{p(\text{B})}{p(\text{G})} \times \frac{p(\text{"B"}|\text{B})}{p(\text{"B"}|\text{G})}$$

(Gigerenzer & Murray, 1987, p. 169)

(The ratio of base rates is the first term on the right; the ratio of the hit rate to the false alarm rate is the second term on the right.) Although Gigerenzer (1991) himself inclines more to frequentism than Bayesianism, his point here is not so much that the Neyman-Pearson solution, in terms of signal detection theory, is the correct one, but rather that no statistical theory has an exclusive claim on modeling human inference and that application of any of them must take into account particulars of the content and context.

Meanwhile, as Gigerenzer considers the earlier findings on Bayesian statistical reasoning from a Neyman-Pearson point of view, his nemesis Tversky has gone on to reinterpret them himself in light of more recent theories. Though Griffin and Tversky (1992) do not explicitly acknowledge Keynes, Cohen, or Shafer in their study of the determinants of confidence, they find their chief explicans in the concept of weight used by these authors. In the simplest example, of coin tosses, the *strength* of the evidence about bias would lie in the preponderance of heads or tails; this quantity can also obviously be assimilated to the earlier concept of representativeness. The *weight* of the evidence, on the other hand, is reflected in the number of tosses. Griffin and Tversky found across a variety of domains that people derive their confidence principally from the strength of the evidence, with insufficient allowance for its weight, and this pattern sufficed to explain the earlier findings both on conservatism and on local representativeness (the “law of small numbers”; Cf. infra).

Clinical Inference and Multiple Regression It is not obvious that the question of multiple regression as a model or standard for clinical prediction falls within the scope of this book, since, unlike the concept of statistical inference, multiple regression makes no necessary use of probability concepts, and, as a model of reasoning, can be considered sheerly as a descriptive technique, as a method of combining quantitative measurements. The concept of statistical inference, however, has evidently carried the curious implication that every statistical technique is a potential model of some cognitive process. Without the support of that concept, multiple regression and clinical prediction might well have remained autonomous developments; as it is, those who contest this assimilation as well find themselves on the defensive.

There are actually two separable issues here. One parallels the question about Bayes’ Theorem or about computer models of cognition in general: whether multiple regression is “wired in” and human inference in clinical contexts simply reflects distortions or imperfections in that process, or whether, at the opposite extreme, there may be no particular relation at all. The second question, though it has aroused greater attention and affect, has little philosophical significance: namely, which method is better in a given application.

Meehl opened the debate in print, as is well known, with his *Clinical and Statistical Prediction* (1954), in which he reported 15 of 16 studies favoring actuarial prediction. Later reviewers criticized his conclusions; Korchin (1976) argued

that most of the comparisons were unfair, inasmuch as the information given to clinicians was typically the same actuarial data that were plugged into the regression equation. The conditions were thus superficially equated, but the result is that clinicians were essentially forced to behave as computers without formulas. When experienced clinicians are given clinical, rather than statistical, data, and are asked to make clinical (individual) rather than actuarial (group) predictions, he argued, their performance surpasses the statisticians'. More recently, Dawes, Faust, and Meehl (1989) have reaffirmed the overwhelming superiority of actuarial methods in their survey and call for psychologists to relinquish their overstuffed egos.

Unquestionably, if the conditions for application are met, there is every reason to expect multiple regression to be superior to human judgment, and no reason not to use it. These conditions include, notably, specification of the relevant variables and routine assessment of large aggregates of individuals, as in college admissions decisions. Sarbin and others have argued that relevance is determined by a shift in probabilities, and all relevant variables must be included in the equation. If, to use Meehl's example, the hallucination of a raven on your husband's head is diagnostically useful, we simply refer to the frequency table for such hallucinations, and doing this formally ensures greater accuracy than doing it implicitly.

The problems with the argument are perhaps apparent. On the one hand, the computer has no basis, apart from the extrinsic one of probability shifts, for assessing relevance, and thus must potentially test everything in the world—an impossible task. This difficulty has been partially obscured by the second, which is that the data—such as statistics on raven hallucinations (this particular kind of raven hallucination)—aren't there in the first place, and it isn't clear they ever will be. It has been a dream of scientists at least since Gavarret (1840) that databases would be compiled to allow treatment decisions to be made on a statistical basis (though Gavarret was limited to binomial distributions, the development of multiple regression being another century away). Such hopes remain largely disappointed, in part because natural populations keep changing (partly in response to changing treatments as well as living conditions). The claims of superiority for actuarial prediction are parasitic on the determination, by clinicians, of what variables to consider for inclusion in the equation.

Whatever practical success actuarial prediction may ultimately enjoy has nevertheless little bearing on whether Gaussian least-squares procedures are or should have been wired in neurologically. It is possible, of course, to take multiple regression as a model simply to study departures from it in human reasoning without necessarily accepting it as a philosophical standard. Attempts to assess the adequacy of such a model, however, still seem regularly frustrated by the complexities of real-life application. As one example, Kahneman and Tversky (1973) fault their subjects for not having the concept of statistical regression as part of their mental program. When graduate students were asked to estimate the 95% upper and lower confidence limits for an IQ score of 140 on "a standard IQ test," only 24 out of 108 students gave regressive intervals (greater probability of true score being less than 140 than greater than 140), while 11 actually gave counterregressive intervals. But these last may well have reasoned the most carefully: for 140 is at, or very near, the upper

limit on some widely used tests, so that someone scoring at the ceiling is quite likely to have a “true” IQ beyond that value.

Causal Inference and Analysis of Variance It is probably inevitable that analysis of variance would have been proposed (Kelley, 1967, 1973) as a model for human inference about causality in social situations, but it is also a little pathetic, just because analysis of variance itself is such a curious instrument for causal inference. The Baconian, or Millian, canons, at least when negatively formulated as principles for narrowing a list of causal suspects (Joseph, 1916), can claim some relevance to causal inquiry in its traditional construction. Textbooks of experimental design may still teach the Baconian logic of holding all factors constant but one, but analysis of variance systematically varies them all at once; then, its main effects represent comparisons averaged across all other factors. Lovie (1979), quoting Underwood’s text as an example, cites this discrepancy as a reason for the slow acceptance of factorial ANOVA among experimental psychologists. The analysis of simple effects, as is urged by some authors (e.g., Marascuilo & Serlin, 1988), accords better with the Baconian strategy, but that is not the most common approach, nor what Kelley had in mind. To be fair to Kelley, he did not develop the ANOVA framework very literally, and he actually sees ANOVA as instantiating the Millian canons, *pace* Lovie.

Like multiple regression, ANOVA can be used descriptively—that is in fact the way Fisher introduced it, as “analysis of variance” more than a technique for inference about mean differences; the use of probabilities is incidental. For this reason, again, the use of ANOVA as a model of thinking falls outside the scope of this book, though it would appear to hold very little promise. The relevance of ANOVA and multiple regression here lies just in illustrating how powerfully the concept of statistical inference has pulled for a more general conception of “cognition as intuitive statistics.”

10.1.4 General Issues

Some of the issues that arise in the study of reasoning under conditions of uncertainty are much the same as those encountered in deductive reasoning. I shall briefly consider here so-called atmosphere effects and the difficulty most people experience with abstract reasoning in general. Issues relating to the perception of randomness, of course, are unique to statistical contexts.

Atmosphere Effects and the Difficulty of Abstract Problems In both deductive and nondeductive reasoning situations, it has been amply shown that human inference is dependent on content and context, the latter including demand characteristics of the laboratory setting.

One specific demonstration was provided by Hammerton (1973), using the Bayesian paradigm; his results were notable as an early example contradicting the

conservatism hypothesis. He gave his participants (mostly housewives) the following statements:

1. A device has been invented for screening a population for a disease known as psylcarapitis.
2. The device is a very good one, but not perfect.
3. If someone is a sufferer, there is a 90% chance that he will be recorded positively.
4. If he is *not* a sufferer, there is still a 1% chance that he will be recorded positively.
5. Roughly 1% of the population has the disease.
6. Mr. Smith has been tested, and the result is positive. (Hammerton, 1973, p. 252)

They were then asked to estimate the probability that Mr. Smith is in fact a sufferer. The probability given by Bayes' formula is .48; the median estimate of participants was .85, with only 1 out of the 10 underestimating. Hammerton also found that deleting some statements made little difference in the results. If one or more from among Statements 3, 4, and 5 were deleted, no numerical answer was possible; but people not only continued to give similar answers: their confidence in their answers was not seriously reduced.

Hammerton considered two possible explanations for his findings. One was the Kahneman and Tversky (1972) hypothesis of the importance of specific over general information, which he took to imply that Statement 3 (or, perhaps equivalently, 2) was decisive. The other was that people have rigid prior beliefs about medical tests; if they regard them as virtually infallible, as they are widely encouraged to do, then the subsequent data of the problem may have little effect. To try to decide between these, Hammerton changed the material of the problem, from a medical test to a screening device for internal cracks in engine parts; the numbers remained the same. Individuals' estimates, however, declined to a median of .60, with 7 out of 20 underestimating. The estimates were still high, leaving room for the Kahneman and Tversky hypothesis, but the substantial difference from the previous results supports the other interpretation, which is similar to the "atmosphere effect" (Woodworth & Sells, 1935) found in syllogistic reasoning.

A variety of more or less extraneous factors affecting probabilistic reasoning can be subsumed under the generalized atmosphere effect. Marks' (1951) and Irwin's (1953) finding that people predict differentially for success and failure could be considered a special case of this general effect. As will be mentioned below in discussion of the ontogenesis of probability, *probably* has emotional connotations which may interfere with reasoning. Similarly, with considerations of justice (Le Bonniec, 1970; Piaget & Inhelder, 1951): People's view of chance as fair ("It's the red one's turn," etc.) distorts their conception of randomness. Children are apt to make probabilistic predictions (predictions in a probabilistic context) on the basis of their favorite color (Goldberg, 1966), but adults are right with them in favoring certain patterns, as was demonstrated in the famous Zenith radio experiment (Goodfellow, 1938). Finally, gambling itself has a certain utility (positive or negative, depending on the person). At least in laboratory experiments where people cannot emerge as actual losers, they tend to prefer long shots to a consistent strategy of maximizing their expected return (Hochauer, 1970; Ross, 1966).

In general, formulation of the concept of atmosphere effects rests on the premise that our reasoning should be similar in problems that are formally similar; the

problem is that judgments of formal similarity often entail some subtleties, and psychologists have been somewhat more likely than their participants to overlook these.

The intrusion of atmosphere effects into problems which experimenters construct as formally similar is not surprising in view of the fact that most people are concrete rather than abstract in their cognitive style (e.g., Keirsey & Bates, 1984) and thus have special difficulty with abstract problems. The Wason selection task (e.g., Johnson-Laird, 1988; Johnson-Laird & Wason, 1970) has generated a large and rich literature on deductive reasoning. Four cards are laid out with their faces displaying, for example, an A, a D, a 4, and a 7, with the information that each card has a letter on one side and a number on the other. People are then asked which two cards must be turned over to test the truth of the rule, “If a card has a vowel on one side, then it has an even number on the other.” Over 90% of people select the A and the 4, evidently owing to a tendency to interpret the conditionality bidirectionally.

Wason and Johnson-Laird have also used this problem to demonstrate the atmosphere effect: if the cards say “Manchester,” “Sheffield,” “car,” “train” and the rule is “Every time I go to Manchester, I travel by train,” fewer than 20% of people get it wrong. Presumably, this problem is easier because the converse is less plausible—“If I travel by train, I go to Manchester.” It might be supposed that letters and numbers are not much less concrete than cities and trains, but the connection in the letter-number problem is still arbitrary and thus lends itself more easily to misinterpretation.

I suspect the tendency to interpret directional implication bidirectionally is closely related to the common confusion over conditional probability. Anyone who has tried to teach conditional probability to psychologists will not be surprised by Maya Bar-Hillel’s (1984) finding that people do not easily distinguish between $p(D|H)$ and $p(H|D)$, and Scholz (1987) observes from his research that many people are unable to formulate conditional probability statements in their own words.³

In general, probability theory introduces additional complications to the difficulties people experience with logic problems. Surely, any reader of the literature on

³Leda Cosmides (1989) recently introduced a fecund program of research on conditional reasoning with her evolutionary social contract theory. She showed that responses on the Wason selection tasks varied according to whether the conditional proposition to be tested could be construed as a social contract (“If you P for me, I will Q for you”). Further research by Gigerenzer and Hug (1992) clarified that people selected cards specifically so as to detect cheaters on the social contract. When the context for the proposition cued participants to the perspective of, say, an anthropologist seeking to understand local customs, the proportion of $P \& \text{not-}Q$ responses (the correct responses according to truth-table logic) was the same as for propositions not expressing social contracts, but if people were cued to the perspective of a party who could be cheated, they selected either $P \& \text{not-}Q$ or $Q \& \text{not-}P$, according to which was required for the detection of cheating. A few participants (notably more in Gigerenzer and Hug’s Konstanz sample than in Cosmides’ sample from Harvard) adhered to the dictates of propositional logic (material implication), but they acknowledged some cognitive strain in doing so when the context clearly pulled for cheater detection.

Bayesian information processing will have been impressed at least occasionally by the complexity of the problems people were asked to solve. Many psychologists, who have been trained in statistics, would still have difficulty solving them with pencil and paper. Cameron Peterson had some appreciation of this long ago; by way of explaining the finding of conservatism, he suggested that

Many of those experiments have been so elaborate that the subjects couldn't be expected to have any idea what on earth their task was. It is easy to see why, with a clear starting position and a confusing task, subjects should not revise their opinions very much. (in Edwards, 1968, p. 18)

One might as well probe, I should think, people's intuitive judgments about the singularity of matrices.

Such is the embarrassing end to which the model of cognition as intuitive statistics appears to lead. On the one hand, people have only limited abilities at manipulation of formal symbols, and ordinary syllogisms are often beyond their grasp. On the other hand, actual human reasoning processes are so complex that cognitive scientists are beginning to despair that they could ever be modeled (H. Gardner, 1985). In the final section of this chapter, I return to consider what may be said of the relation between logic and statistics, on the one hand, and reasoning, on the other.

Judgments of Randomness Randomness, in the secondary sense defined by Spencer Brown (1957), is the absence of discernible patterns (Chap. 6). The ability to discern patterns in data, however, is a point of some pride, as well as a survival skill. Perhaps, for this reason, we are not good at recognizing or constructing random patterns. Put another way, we are too quick to interpret and leave too little to chance. If people are asked to generate random strings of binary digits with an overall proportion of 1:1, they violate normative theory by making small units (samples) conform to the 1:1 (population) rule. In other words, they make too many alternations and leave too few runs. This observation was made informally by Borel (1939/1952) and Reichenbach (1949); it has been demonstrated experimentally by Ross (1955, 1966), Ross and Levy (1958), Gratch (1959), Bakan (1960), and John Cohen (1973).

The expectation that small samples will follow the rule for the whole series is at the root of the so-called gambler's fallacy. It is also known as the fallacy of the maturity of the chances and as the negative recency effect, though L. J. Cohen (1977) suggests that the terms be distinguished, with "gambler's fallacy" reserved for actual gambling situations. The negative recency effect is the prediction, following a run, of the nonrun event, essentially the expectation that "the law of averages" must prevail over the *short* run. Negative recency is an adult achievement; Ross and Levy (1958) showed that it is not acquired before about the age of 15.

It is also the object of considerable scorn by many writers on probability, but it may be another case where unsophisticated intuition has as much plausibility as formal theory. It is, for one thing, supported to some extent by everyday experience. In many naturally occurring series, we never allow runs to become too long before we intervene to right the balance; casinos, for example, daily change roulette wheels

to prevent the wearing of the mechanism in any regular way. But more important is the argument made by Spencer Brown (1957) and L. J. Cohen (1982), and by d'Alembert (1767) long before them. Suppose we want to decide whether a roulette wheel is biased, and we adopt as a criterion of significance a run of 20 reds or 20 blacks. Suppose then that at some point, we get a run of 19 reds. Our next bet simply expresses our belief about whether the machine is biased or fair. A bet on red implies a judgment that it is biased; but this is not necessarily any more or less defensible than a bet on black, implying a judgment that the wheel is a good randomizer. The point is only that if we are committed to the belief that the wheel is fair, we may be committed to a maturity of the chances betting policy, but the stigma does not necessarily go any further.

L. J. Cohen (1982) adds the observation that, if we attempt to account for the gambler's fallacy in terms of a heuristic of representativeness, we have trouble explaining why people predict the nonrun event only after a long run and not after a short one. To predict a red after a single black, or two reds after two blacks, would be even more representative. But the general problem of the gambler's fallacy disappears if we recognize, with d'Alembert and Spencer Brown, that only in the hypothetical world of formal mathematics do we utterly rule out the possibility of bias.

10.1.5 Randomness, Representativeness, and Replication in Psychological Research

The tendency to see things as patterned and meaningful rather than random has special consequences for psychologists, since the assumption of randomness is the basis of our traditional methodology. As was noted in Chap. 9, many textbooks explicitly confuse randomness and representativeness. Smith (1934) was cited as a historical example in Chap. 9, but modern textbooks show scarcely any improvement in their understanding. A more recent instance is the generally excellent textbook by Marascuilo and Serlin (1988): “The basic unit is called the *population* or *universe* and includes all people, objects, and concepts for which the subpopulation in the experiment, called a *sample*, can be considered representative” (pp. 11–12). Cook and Campbell (1979) are careful to include reference to sampling variability, but when, in characterizing their “random sampling for representativeness model,” they say, “Formally speaking, the most representative samples will be those that are randomly chosen from the population” (p. 75), the concepts of randomness and representativeness are made to sound closer than they are. We should properly hesitate to take any “sample” of 3 tosses of a coin as representative of the infinite “population” of tosses that might be observed with the coin, and the same would be true with any sample of 3 million, insofar as 3 million is still an infinitesimal fraction of the possible observations. With a finite population, on the other hand, such as the voters of Calaveras County, larger samples do become more representative and eventually exhaustive. But psychologists say they are making inferences to infinite

populations on the basis of small samples; and, if we are speaking of randomness in the primary, process sense (Chap. 6), we can expect random samples to be unrepresentative with a frequency inversely related to their size.

Three or four generations ago, there would have been some justification for the confusion over these concepts, when samples consisted of many hundreds or thousands of observations and the object was really estimation of population parameters: Random samples would tend indeed to be representative. The shift from estimation to significance testing, coinciding with the shift from large to small samples, brought the concept of randomness to the fore and put the burden of inference directly on the sampling distribution. As was discussed in Chap. 9, psychology-size samples are much too variable to treat as representative—except that we obscure this variability from ourselves by never taking more than a single sample. If large numbers of replications were commonplace, we might have a chance to observe the variability of our results, and the practice of significance testing would be severely threatened. Instead, we expect that individual samples should be representative of the population, and therefore homogeneous, and hence that results of significance tests should be much more replicable than they are—hence, the “replication crisis” discussed in Chap. 1. The final link in the implicative chain is the systematic overestimation of power in psychological experiments. Tversky and Kahneman (1971) refer to the phenomenon as the “law of small numbers”—the belief in “local” representativeness, or the assumption that the law of large numbers should be manifested at the local level as well as the universal.

Oakes (1986) offers a simpler explanation of psychologists’ belief in the replicability of their results, not necessarily inconsistent with the representativeness hypothesis. If, as the data of his that I presented in Chap. 1 showed, most psychologists endorse an inverse interpretation of significance levels, then to them a significant result means that the alternative hypothesis has a high probability of being true, and hence the effect should be manifested in replications. Oakes attempted to drive a wedge between his “significance hypothesis” and the representativeness hypothesis with variations on the problem posed by Tversky and Kahneman (1971).

- A. Suppose you are interested in training subjects on a task and predict an improvement over a previously determined control mean. Suppose the result is $Z = 2.33$ (sig. $p = 0.01$, one-tailed), $N = 40$. This experiment is theoretically important, and you decide to repeat it with 20 new subjects. What do you think the probability is that these 20 subjects, taken by themselves, will yield a one-tailed significant result at the $p < .05$ level?
- B. Suppose you are interested in training subjects on a task and predict an improvement over a previously determined control mean. Suppose the result is $Z = 1.16$ (NS $p = 0.12$, one-tailed), $N = 20$. This experiment is theoretically important, and you decide to repeat it with 40 new subjects. What do you think the probability is that these 40 subjects, taken by themselves, will yield a one-tailed significant result at the $p < 0.05$ level?
- C. Suppose you are interested in training subjects on a task and predict an improvement over a previously determined control mean. Suppose the result is $Z = 1.64$ (sig. $p = 0.05$,

one-tailed), $N = 20$. This experiment is theoretically important, and you decide to repeat it with 40 new subjects. What do you think the probability is that these 40 subjects, taken by themselves, will yield a one-tailed significant result at the $p < 0.05$ [sic] level? (Oakes, 1986, pp. 83–84).

The key question here is (C), but Oakes's argument is difficult to follow because of a misprint in that question; the question about the 40 subjects should have asked about significance at the 0.01 level. In that case, he argues, people under the representativeness hypothesis would expect the second result to be too close to the first, and hence not significant at the 0.01 level, whereas under the significance hypothesis, regarding the effect as established, they would expect it to be replicated. For questions (A) and (B), the two hypotheses yield similar predictions: a high probability in (A) and a low probability in (B). The true probability in all three cases, construed as the power of the second test against the effect size found in the first, is .5. The mean responses to the questions—.798, .292, and .747, respectively—confirmed Oakes's explanation in terms of the significance hypothesis.

Oakes's experiment is vulnerable to the same criticism as many others probing intuitions about probability, that the problems are very difficult to solve in one's head. His subjects, however, were psychologists who should have known how to solve them at least on paper. They were in fact offered pencil and paper, but few had any idea what to do, and some denied that any precise calculations were possible.

There are several unfortunate consequences of psychologists' belief in the law of small numbers or in the invertibility of significance levels. (a) Sample size has usually been dictated by expediency rather than advance considerations of power; the law of small numbers lulls us into believing that our experiments are more powerful than they are. Jacob Cohen (1962) estimated the power of the average psychological experiment (in the *Journal of Abnormal and Social Psychology*) against a medium-size effect to be .48; in other words, the investigators were giving themselves less than an even chance of detecting true medium-size effects. Thirty years later, Sedlmeier and Gigerenzer (1989) estimated the average power of studies in the same journal to be just as poor as ever, thanks to the practice of adjusting alpha levels for multiple comparisons, to which psychologists' attention was drawn at about the same time (Ryan, 1959). (b) Replicability, for us, means not just results in the same direction; they have to be significant as well. Overestimating the probability of replication, we are inclined to respond to marginally significant results by repeating the experiment—or making our students repeat it—when there is precious little hope for successful “replication.” (c) By focusing on individual samples and expecting representativeness there, we attach undue confidence to early trends and to the stability of observed patterns. If we actually repeated experiments large numbers of times, we would get quite a different picture from what we see by looking at only the first member of the series. (d) We underestimate sampling variability and hence the width of confidence intervals. Indeed, we may deny the reality of sampling variability altogether, as we do in seeking a theoretical explanation for every discrepancy between repetitions of an experiment. As Tversky and Kahneman

(1971) observe, this practice deprives us of opportunities for observing sampling variability in action and thus reinforces belief in the law of small numbers.

10.2 The Ontogenesis of Probability

By way of trying to gain a more detailed picture of how probability concepts develop, I once interviewed nine children aged 8–12 years. As an empirical study, it is rather severely limited, but the results do no violence to everyday observations on the subject, and they can still serve to illustrate the struggles we may all have gone through with these concepts.

The adverbial form *probably* is normally acquired first, by around the age of 9. It is used initially to mark a state short of certitude but sometimes also to express a hope or a strain at confidence. One youngster of 11 said that in saying “probably,” “You’re just trying to be confident.” Another, age 10, acknowledged that in saying, “I probably will not have any cavities when I go to the dentist next time,” she was expressing a hope. The use of *probably* as a hedge, interestingly, is experienced by some children as indicating almost the opposite of its usual meaning. Thus, a boy of 11, when asked what *probably* means, said, “It’s usually a lie”; by way of explanation, his example was asking his father whether they could go to the beach that day, and his father’s response was “Probably.” Clearly, at this point, in children’s use and understanding of *probably*, the concept does not serve merely as a simple, objective appraisal of evidence; any epistemic meaning it carries is thoroughly suffused with emotive intent.

The next step in the evolution of the concept focuses more on its epistemic aspect, especially in comparison with other qualifiers of certitude, like *maybe* and *might*, which are typically offered by young children as synonyms of *probably*. When queried about comparisons among these, children sometimes spontaneously construct a dimension of sureness, with *probably* toward the upper end, which is “positive.” It is notable that, for one reason or another, a number of children place *probably* near certainty. Some of them, for example, are able to assign numbers, on a scale of their choosing, to show roughly the relative positions of the various concepts they include on the dimension, and on a scale of 100, 99 is a popular choice for *probably*, with *maybe* and *might* commonly assigned values in the 40–60 range. As L. J. Cohen would predict, the lower end of the scale is ambiguously defined; children are confused about whether it represents maximum certainty that not-X or merely a lack of evidence supporting X. When they acquire a concept of relative frequency with which to match their scale of confidence, then they have to forget that the latter scale ever made any sense.

Mathematical probabilities being proportions (in the finite case), comparisons among them involve relations among relations and hence, according to Piaget and Inhelder (1951), cannot be mastered until the formal operational stage, around 12–15 years. Some of Piaget’s findings have been questioned by American psychologists claiming to have demonstrated probability judgments or concepts in

children as young as 4 (e.g., Davies, 1965; Goldberg, 1966; Messick & Solley, 1957), but Hoemann and Ross (1971) criticize these findings in turn on the ground that not all the tasks employed were adequate to justify the claims. They point out that “successfully choosing the favorable odds need not give an index of probability knowledge” (Hoemann & Ross, 1971, p. 233) and that “not all tasks which are nominally probability tasks require use of probability concepts” (p. 234). These authors contend that the problems posed in such studies were essentially a matter of comparative magnitude estimation, which is more reasonably within the reach of preschoolers. In their own experiments, they presented children with, for example, a spinner with black and white sectors; then, the *probability* task was a question about which color the spinner would come to rest on; the *proportionality* task, by contrast, was simply the question whether there was more black or white on the disk. Though the two tasks are equivalent in difficulty for children older than 8 or 10 years, younger children (4–8 years) did significantly better on the proportionality task, which Hoemann and Ross argue is nearer the task used by Piaget’s critics. Their experiment is interesting because their distinction between probability and proportionality corresponds, for continuous sets (the task could have been modified, of course, to involve discrete probabilities), to the distinction between probability and relative frequency and their results testify to the psychological reality of the distinction.

The word *probability* itself is typically acquired around the age of 12 or 13, which is just when children are also acquiring the formal operations necessary for grasping the principles of the frequency calculus—permutations, combinations, and so forth. The word *probability* from the start is given a mathematical meaning, and the lexemic similarity to *probably* (and to *probable*, which seems to appear even a little later than *probability*, presumably because of its comparative rarity in ordinary speech) suggests a more substantial connection. We may see rather nicely, from the psychological point of view, some of what is involved in coordinating these two concepts by watching the struggle of a precocious 12-year-old reflecting on the problem for the first time.

Asked about the words *probably*, *probable*, and *probability*, he began by defining *probably*, not simply as a modal term, as younger children do, but in terms of odds:

Probably I use as a figure of speech to mean most likely, like 40/60 odds, and the 60 is good.
You have a better chance if something is probable, and when you’re using it in a sentence,
as an adjective, you say *probably*.

He went on to compare *probably* with *maybe*, which he saw as “less definite.”⁴ Interestingly, he conceived the uncertainty in *maybe* as compound, so that *maybe*, instead of being restricted to a middling range, “could be 99/1 or 70/30 or anything; it’s just an uncertain chance, like a big question mark.” The two numbers had to add up to 100, he implied, because they were percentages. When asked what they were percentages of, he construed them straightforwardly as objective relative frequencies: a

⁴The notion of precision seems close to the idea of definiteness articulated here; compare the response of an 11-year-old who said that *probably* is “more exact” than *maybe*.

60% chance of rain, in his example, meant that out of 100 occasions of similar weather conditions, it would rain on about 60. The variability—which he spontaneously acknowledged—he attributed to imprecision in the weather forecaster's measurements; the ambiguous implication is that “complete” information would lead to a precise prediction of 60% (or a similar figure).⁵

Now, for the test of his convictions: when questioned about the application of the frequency interpretation to the single case—on the meaning of “60 times out of 100” when applied to the forecast for a particular day—he responded at once by partitioning a single day into 100 time periods, or microevents, so that a 60% probability of rain for the day as a whole would translate into a prediction of rain for 60 of those time periods. He was not strictly speaking reductionistically, since “There’s actually two meanings you could get”: “60% of the day it could rain, or there’s 60% chance in the way the sky looks that it will rain.” Ingenious as it was, his solution worked less well in other examples: the toss of a coin could less easily be divided into 100 parts than the span of a day. “You would need at least two tosses to be sure of 50/50 odds, because you can’t toss half a coin”; but even then, the coin was not constrained to fall once heads and once tails, because, he explained, the odds stayed 50/50 from toss to toss. In the space of two interviews, he never succeeded, of course, in working out a solution even to his own satisfaction, but his attempt to give meaning to singleton probabilities by constructing reference classes is, in rudimentary form, the same option elected by frequentists like Reichenbach.

It is interesting to juxtapose Piaget and Inhelder’s (1951) discussion of the same problem, partly because they take issue with Borel (1939/1952) in a debate in which both parties appeal to psychology for support. Borel argues that

The probability of an isolated case is the basis of the probability calculus. This notion is natural to each of us, just as the notions of hot and cold; but experience and reflection allow us to make it precise and to obtain rather close approximations of probabilities, just as we can succeed well enough in estimating, without a thermometer, the temperature of the air in which we live. (Borel, 1939/1952, p. 104)

Some cases may involve purely rational estimation: Few of us have any experience at all with regular icosahedra, but we would scarcely hesitate to attribute a probability of 1/20 to any particular face falling uppermost.⁶ In other cases, we may have some experience, but too little to speak plausibly of a class of events. He gives the example of assigning a probability to the proposition that the weight of an ordinary suitcase falls within certain limits.

Piaget and Inhelder (1951) dissent entirely.

⁵The same confusion crops up occasionally in the philosophical literature on probability (in Reichenbach, for example), in the assumption that more and more evidence would impel us inevitably toward a probability of 1 or 0.

⁶The interview data described above provide some small amount of psychological support for Borel’s argument from the icosahedron: for precisely this example was spontaneously offered by the 12-year-old whose theorizing was just described.

From the psychological point of view, it seems very difficult not to see in every probability judgment a reference, explicit or implicit, to a distribution or frequency system. Indeed, we have established that the intuition of probability was not at all primitive for the child and that it emerged only with this consideration of the whole system. (p. 260)

They admit that children even younger than 7 or 8 years sometimes make subjective, nonfrequency judgments of probabilities of single cases, but they discount these simply because they do not proceed from an understanding of combinatorial chance.

As for the weight of a suitcase, it is difficult to believe that an individual confident of the limits >1 kg. and <50 kg. would never have had the occasion of making previous experiments in this situation: a frequency distribution is thus surely inscribed in his or her practical habits, even if it contains only quantifications such as often, ordinarily, rarely, almost always or almost never, etc. (pp. 260–261)

They conclude that “The elaboration of a system of distributions is then the precondition of probabilistic intuitions” (p. 260).

One has the sense that the debate here, as perhaps in the history of probability in general, is taking place on a more stipulative or definitional than substantive level; it has, at any rate, a lack of conclusiveness appropriate to such a debate. Piaget and Inhelder (1951) oversimplify the philosophical situation seriously in claiming that

Certain axiomatic or logical works on the probability calculus, such as those of von Mises or Reichenbach, have served to show that every probability is necessarily relative to a collection of events, hence to a certain distribution or to a system of frequencies given from the start. (p. 260)

They essentially restrict the meaning of probability to mathematical probabilities, and from this perspective reject any other uses as nongenuine. The appeal to psychological data scarcely helps either party. Borel is right in arguing that children have a primitive intuition of probabilities not based on relative frequencies, and Piaget and Inhelder are right in denying that whatever intuitions they have are based on an apprehension of combinatorics.

The foregoing sketch of the development of probability concepts can be seen as providing a modest illustration of Piaget’s controversial thesis that ontogeny recapitulates history (Wartofksy, 1971). Both developmental sequences start from a nonnumerical evidential concept, then confront a lexemic relative referring to relative frequencies. The latter resolves the ambiguity in the lower bound of the scale but introduces an ambiguity in the meaning of singleton probabilities. Those who grow up to be philosophers then set themselves the task of making sense of a dualistic concept—attempting remarkably seldom, however, to get beyond the philosophical context in which the concept was formed.

10.3 The Propensity Theory of Probability

The propensity theory is historically the most recent of the major interpretations of probability, and in the second half of the twentieth century, it supplanted the frequency theory as the most discussed approach in the literature. Like the dualistic interpretation, it seems often to be implied in everyday usage and so on that ground could conceivably claim some very remote antecedents, but actual talk of probabilities as propensities appears to be a phenomenon of the late nineteenth century.

Interestingly, some of the principal figures in this line of thought are associated with frequentism. One of these is Peirce, who, in notes appended in 1910 to his papers of 1878 (Peirce, 1932, §§664–667), spoke of the “would-be” of a die and compared that property to ordinary habits of humans. The would-be of a die to turn up a 6 or a 3, Peirce said, finds its expression in the long-run relative frequency of these numbers among the tosses, but is not identical with their relative frequency. Peirce never developed the implications of the propensity view, however, and it remained undeveloped for nearly 50 years.

The first to announce a propensity theory of probability under that name was another well-known frequentist—Karl Popper (1957/1962, 1959). Popper originally became interested in probability theory in connection with the problems of quantum physics. He was particularly concerned to eliminate “the intrusion of the subject into physics” (Popper, 1957/1962, p. 65). Heisenberg had introduced the observer formally into physics through the Uncertainty Principle, and the determinist view required that a subjective interpretation be placed on the probabilistic statements of quantum theory, inasmuch as uncertainty could only be located in the observer. Popper rejected the subjectivist interpretation, which “drags in our knowledge” (1957/1962, p. 69), in favor of the frequency view, which was the only objective interpretation he saw at the time. It was a particular problem with the frequency view, however, which led him into the propensity interpretation.

Imagine, he says, a long sequence of throws with a loaded die, by means of which we have established that the probability of a 6 with this die is $\frac{1}{4}$. Now consider such a sequence, into which have been inserted two or three throws of a fair die. Then, Popper argues, we shall have to say that the probability of a 6 on these throws is $\frac{1}{6}$, “in spite of the fact that these throws are, according to our assumptions, *members of a sequence* of throws with the statistical frequency $\frac{1}{4}$ ” (Popper, 1959, p. 32). In considering possible responses to this paradox, Popper rejects appeals to special knowledge of these throws as subjectivism; we can, in any case, make it harder for the subjectivist by stipulating that we do not know which of the many throws were made with the fair die, so that the reasonable betting quotient is still (nearly) $\frac{1}{4}$. He likewise rejects the response of the frequentist, which must refer to long sequences with the fair die, for the reason that we have only two or three throws with it, which cannot possibly exhibit a relative frequency for 6 of $\frac{1}{6}$. Popper thinks this example forces frequentists to accept as admissible sequences in their theory only (virtual or actual) sequences which are “*characterized by a set of*

generating conditions" (1959, p. 34); then, the sequence given in the problem is no longer an admissible reference sequence for probability statements.

But this modification amounts to a shift to a propensity view of probability. For probability is now taken to refer to a property, not of an actual sequence itself, but of the generating conditions of a sequence—"a tendency, or disposition, or propensity to produce sequences whose frequencies are equal to the probabilities" (1959, p. 35). Popper stresses that the propensity is a property of the whole setup—not just of the die itself, but of the die together with the manner of its being thrown. He also indicates how propensity, on this view, is to be distinguished from possibility, which is a related notion long associated with probability: The difference is that mere possibility carries no tendency to realize itself, whereas it is precisely that disposition to which Popper wishes to call attention in probability. Finally, probability remains objective on the propensity interpretation, as it is on the frequency view. The observer does not enter at all; what interferes is only a change of experimental arrangements (1957/1962, p. 69) (insofar, presumably, as arrangement and arranger can be logically distinguished). Unlike frequency probabilities, however, propensity probabilities can also be applied to the single case, measurable by reference to a *potential* statistical frequency rather than an actual one. To speak of the probability of a 6 as $\frac{1}{6}$ is to advert to the tendency of a 6 to realize itself on that particular throw, but evidently, only the sequence can tell us what that propensity is.

Hacking (1965) has espoused a similar conception. He agrees with Popper that propensity is a property of the "chance setup" and that it pertains not to individual trials, but to trials of a designated kind. One of his reasons for insisting that propensity and therefore probability is relative to a kind of trial has to do with the classical frequentist argument: If on a given toss a coin falls heads, then its *true* propensity, if only we had known it, was $P(H) = 1$. Hence, the true chance of heads on any individual toss must be 0 or 1. Hacking's reply is to ask on what kind of trial it is alleged that $P(H) = 1$. It could only be tosses in which exactly the same causal conditions were operative as in the given trial; this kind of trial can be identified, in principle, and true enough, the probability of heads for tosses in this class is 1. But this is simply not the kind of trial we are ordinarily interested in; rather, we ask about the kind of trial when we merely toss the coin and note the result. A different probability obtains for this kind, though it is not inconsistent with the first.

Mellor (1971), on the other hand, disagrees with both Hacking and Popper and joins Peirce in taking propensity to be a property of the individual object—the die or coin, for example—rather than a property of the entire experimental arrangement. To follow the latter course, he suggests,

would be like saying that a thin glass together with a hard stone floor might be fragile, or that a grain of salt with a bucket of water might be soluble, or that a fire together with a thermometer might be hot. (Mellor, 1971, p. 75)

In truth, there is some ambiguity in the concept of "experimental arrangement" (putting aside the fact that we may want to speak of probability and chance in non-experimental contexts, for instance in mortality statistics). What must be held constant across trials for us to speak of an experimental arrangement or a set of

generating conditions? As Kyburg (1974) observes in his review of propensity theories,

It is not immediately clear. We may surely consider throwing the die on various surfaces; we may consider having it thrown on various days of the week, during various seasons, at various times of day. We may consider throwing it by hand, or with a dice cup, or by means of a specially designed machine. We may consider performing the experiment ourselves, or having someone else perform it, or having a number of different people, of different ages, of both sexes, of varying ethnic origins, perform it. It is difficult to see what is held “constant” except for the die, and the throwing of it. But this is too generous: for we can surely arrange for the die to be thrown by a machine—a very carefully constructed machine—in such a way that it always (or almost always) lands with the six up. It begins to seem as if what is held constant is just those circumstances that determine the chance distribution of outcomes—but of course, that would not do, on pain of circularity. (p. 360)

Kyburg's criticism strikes at the core of the propensity theory, and it is not immediately clear what sort of response is possible. The most general point to be made is that the specification of the system exhibiting the propensity in question must be made on external, or extrastatistical, grounds and furthermore that it may well not be possible to formulate any general description. This, at any rate, is Gière's (1976) answer, and we may concretize his argument with the example of so-called mixed, or randomized, tests. Mixed tests present essentially the same problems as Popper's mixed sequence with dice, and they provide a stock counterexample to all sorts of statistical claims. It will be sufficient to refer to a specific example.

In sampling from, say, a standard normal population, we decide to take a sample of size 10 if a fair coin comes up heads, a sample of size 100 if it comes up tails; a replication of the experiment will always begin with a coin toss. Hence, for a one-tailed test at the 5% level, the cutoff will lie somewhere between $1.645(10^{-1/2})$ and $1.645(100^{-1/2})$. But if we know the sample size is 10, what justification do we have for taking advantage of the superior characteristics of the mixed test? Or if we know the sample size is 100, why should we in effect throw away data, in using the mixed test? The main reason is that to do otherwise is to make our analysis conditional on the outcome of the experiment. If we are prepared to do that, there is nothing in principle to stop us from conditioning the analysis on all our data, giving them always a probability of 1.

The mixed test is appropriate in situations where we have a stooge who performs the experiment without letting us know the result of the randomizer (the coin, in this case). In other cases, Gière thinks we are justified in using the cutoff appropriate to the particular sample size obtained, and our justification depends on our specifying which system is operative in our experiment. If we have a stooge, then the system necessarily includes the randomizer, and generates the mixed sequence; but otherwise, the system may be defined with reference to a particular sample size, and the propensity will be a property of that (more restricted) system, and the appropriate sequence will be unmixed. But the general point is that it is up to us to define the operative system in a theoretically meaningful way, and there are no formal principles which are in themselves adequate for this task.

Mellor (1971) makes a related point about specification of the system exhibiting the propensity in question: namely, that it is a matter of convention which element of the system is selected to bear the disposition. If a glass breaks on falling to a stone floor, do we attribute the break to the fragility of the glass or the hardness of the floor?

It depends on whether relevantly similar glasses break on relevantly different floors and whether relevantly different glasses break on relevantly similar floors. Without an array of other properties of glasses and floors and a network of laws to tell us which are relevant to fragility and hardness, we cannot settle the question. It is the same with propensities. (p. 90)

The choice, though conventional, is not arbitrary, for we must identify propensities in such a way as to tie into a network of laws, so that we gain some explanatory power beyond the immediate event the propensity was designated to explain.

There is one further aspect to the issue of where propensities are to be located. Do we locate them in reality, making propensity an objective property of things, or do we make them an epistemological concept, relative to our knowledge? Most propensity theorists have spoken as if propensities were physical properties of causally indeterministic systems or processes and were *manifested* in individual trials or cases. A theory of probability is thus made possible which is both objective and applicable to individual events. But as Gière (1976) acknowledges, instances of absolute, single-case propensities are rare; the only present contender is quantum physics, and the indeterministic interpretation is controversial even there.⁷ There is no compelling reason, however, for making the ascription of propensities to objective reality; for the purposes of probability theory, a relativized conception may suffice. Its principal drawback, as Gière notes, is that it leaves open the question of the meaningfulness of attributing propensities to individual trials.

We are close here to what is surely the major problem with propensity theories, in the eyes of contemporary appraisers. It is the same objection that can be raised against any dispositional concept: that it pertains to unobservable, hypothetical aspects of entities and adds nothing to a theoretical account that is not present in a description in terms of observable features. Popper (1959) has directly addressed this issue, and fittingly so, since he spoke so disparagingly of the propensity interpretation in the original edition of *The Logic of Scientific Discovery*.⁸ He boldly accepts the characterization of propensities as unobservable physical properties and invites comparison with the physical concept of force or a field of forces. Like other propensities or dispositions, the concept of force once had anthropomorphic and metaphysical import; it gained scientific usefulness only when stripped of those implications. But the concept of a force field is also an unobservable dispositional

⁷ Among the other advantages of the propensity theory, Popper (1957/1962) boasts that it eliminates the problem of wave-particle dualism which bedevils modern physics, but if he is correct in his claim, no one seems to have noticed or cared then or since.

⁸ The change of mind occasioned some irony. To his original condemnation, he subsequently (1968) added the remark: "This somewhat disparaging characterization fits perfectly my own views which I now submit to a discussion in the "Metaphysical Epilogue" of my *Postscript*, under the name of "the propensity interpretation of probability" (footnote *4, p. 212).

property of a certain physical arrangement, and it has regularly been found useful in explaining observable phenomena of motion. Popper asks basically that we grant the same status to probability.

Mellor (1971) notes, for his part, though, that a probability is admittedly an odd sort of propensity. To say of a glass that it is fragile is to make a subjunctive conditional statement that it *would* break—invariably—if (“suitably”) dropped. If it failed to break, we would either say that its fragile disposition had changed or that we had been wrong about it all along. The result of a chance trial, however, must not be invariable. Mellor resolves the dilemma by declaring that the result of a chance trial is not the “display” of the propensity in question. Rather, that display is found in the *distribution* of chance over all possible outcomes.

What the propensity theory accomplishes is primarily a removal of some of the problems with the reductionist definition of the frequency theory, including the need to specify a series as *definiens*, and in achieving a better match with the everyday meaning of probability, at least for aleatory probabilities. The propensity theory would still appear to be applicable only to aleatory rather than to epistemic probabilities—it is dubiously meaningful to speak of my propensity for being elected President—but for these, it facilitates the link to epistemic concepts like expectation and support. In this respect, it provides a nexus: We expect an outcome because it has a disposition to occur, and the disposition determines its relative frequency of occurrence.

The implications of the propensity theory for statistical inference would thus appear primarily in the possibility of recasting Neyman-Pearson theory in a way that gave meaning to individual case probabilities. One could then speak of the propensity of a given kind of trial to produce a certain result, and significance levels would not need to be interpreted in reductionist frequency terms. It would less clearly license probabilistic locutions about population parameters, such as fiducial probability statements or confidence intervals with numerical limits.

Gièvre (1976) has proposed just such an interpretation and has been criticized (cf. Johnstone, 1988) chiefly on the issue of ambiguity of the reference set in the Behrens-Fisher problem. This was Fisher’s principal exhibit in the case against the frequency interpretation of significance levels; his solution involved a fiducial distribution for the variance ratio which could not be duplicated in Neyman-Pearson theory. The force of this academic example seems to me insufficient, however, to maintain a sharp split between frequency and propensity interpretations.

10.4 The Likelihood Theory of Statistical Inference

The likelihood approach is not new. As was mentioned in Chap. 7, the idea goes back to Lambert and Daniel Bernoulli. The name is due to Fisher, but his advocacy of it as a general method was obscured by the success of significance testing. Neyman and Pearson (1928a; 1928b; 1933) considered likelihoods in the development of their theory, but they rejected likelihood in itself as a criterion in favor of

the p integral, partly because they wanted a test that provided information about errors in repeated sampling. In recent years, Barnard (1949), Birnbaum (1962), Hacking (1965), and Edwards (1972), among others—likelihoodlums, as Chatfield (1972) calls them—have contributed to the theory. I shall concentrate here on Edwards, partly because of the high state of development of his theory, and partly because of the interest of his examples.

Whatever its use in statistical inference, the concept of likelihood cannot be considered logically primitive, for it is just a case of direct probability: the probability of the observations given some hypothesis. The situation is confused somewhat by the fact that the likelihood is then often said to be the likelihood of the hypothesis, rather than of the data, and even Hacking (1972) speaks of likelihood as “a sort of inverse of physical probability” (p. 133).

The “likelihood theory of statistical inference” really amounts only to saying that likelihood, rather than either p integrals or posterior probabilities, ought to be the end product in statistical inference. More precisely, what is advocated as a criterion is the *likelihood ratio*, the likelihood of the data under the hypothesis tested, relative to their likelihood under some contemplated alternative.

Edwards (1972) adds the concept of the *support function*, which he defines as the natural logarithm of the likelihood function. The main reason for taking logarithms is to get an additive measure of support: Since the likelihoods of independent observations are multiplied, the support offered to a hypothesis by one set of data can be added to the support given the same hypothesis by a different set of data. Natural logarithms are of course convenient in working with the exponential family of distributions.⁹

Tests of hypotheses in Edwards’ theory may then be performed in terms of either likelihood or support. For a likelihood test, we simply calculate the likelihood ratio based on the hypothesis tested and the alternative envisioned; a large likelihood ratio implies strong support for the hypothesis. There are no tables or criteria for evaluating likelihood ratios any further, and Edwards claims this as an advantage; essentially the idea is that once we have worked for a while with likelihood ratios, we will acquire a feel for the amount of support implied by various magnitudes. This argument is similar to that given by de Finetti and Savage regarding personal probabilities: that however artificial it is to us now to assign numerical probabilities to hypotheses *a priori*, they will come to seem natural through extended use, just as we have acquired a feel for temperatures on the Fahrenheit scale. Hacking (1972) is still worried that likelihood ratios, unlike p values, carry no guarantee of comparability; we have no formal grounds for assuming that a given value for a likelihood ratio means the same thing in two unrelated experiments.

⁹The definition of support in terms of natural logarithms also links it with the concept of information. Edwards defines the *expected support function* as the mean value of the support function conditional on the true value of θ being θ^* ; then, the *expected information* is minus the curvature of the expected support function at the point $\theta = \theta^*$, where the expected support is maximum. Edwards’ definition of information agrees with that given by Fisher (1925), who further related it to entropy.

A slightly more familiar form of test is obtained by casting it in terms of the support function. Here we may compute the increase in support to be gained by accepting an alternative hypothesis. We may even define “*m*-unit support limits” to parallel confidence limits; for a normal distribution, the 2-unit support limits correspond to ± 2 standard deviations, which are also approximately the 2.5% points. Edwards does not want to be rigid, however, about any cut-offs in terms of support limits.

The Method of Support, as Edwards calls it (to keep it distinguished from the Method of Maximum Likelihood) can be profitably illustrated by the chi-square test for the mean and variance of a normal distribution. For the test of the mean, based on a single observation x , the likelihood function for μ is simply the normal density; hence the support function, down to a constant, is $-(x - \mu)^2/2\sigma^2$. The gain in support to be had by accepting the alternative $\mu = x$ is thus $(x - \mu)^2/2\sigma^2 = 1/2\chi^2$. This increase in support, m , takes the value 2 when $\mu = x \pm 2\sigma$; hence the 2-unit support limits. Now if it is the variance that is to be tested, the support function is $-\frac{1}{2} \ln \sigma^2 - \frac{1}{2}(x - \mu)^2/\sigma^2$, which attains a maximum at $\sigma^2 = (x - \mu)^2$; the increase in support is then $-\frac{1}{2} \ln (x - \mu)^2 - \frac{1}{2} + \frac{1}{2}(x - \mu)/\sigma^2 = -\frac{1}{2} \ln \chi^2 - \frac{1}{2} + \frac{1}{2}\chi^2$. The 2-unit support limits can be found by setting this expression equal to 2; they are 0.0068 and 6.94, which are approximately the 95% and 1% points of the chi-square distribution, respectively. The implications of Edwards’ analysis are especially interesting for the traditional chi-square goodness-of-fit test (see pp. 184–185).

We should note one problem, however, which Edwards insists is only “technical.” When we have only one observation, as we did above, then the best-supported value for the variance is obviously zero, and in fact, we could get an infinite increase in support by simultaneously letting $\mu = x$ and $\sigma^2 = 0$. Why not? Edwards’ answer is that we always have prior reasons for regarding the hypothesis of zero variance as ridiculous. Hacking (1972) is not so easily convinced: If we decide in advance that w is the least possible value for the variance we are prepared to accept, then on that assumption w becomes the best-supported value, and it is still not the value given by the likelihood theory.

The subject of prior support is in fact one of the most curious aspects of Edwards’ theory. The need for such a concept is evidently the same as the need for a prior distribution in Bayesian statistics: The first time we collect a set of data bearing on a hypothesis, we cannot evaluate the gain in support provided by the data without some initial support for reference. Edwards achieves this initial support by the device of imaginary experiments (which is also used by Good, 1950, in obtaining prior personal probabilities):

The *prior support* for one hypothesis against another is S if, prior to any experiment, I support the one against the other to the same degree as if I had conducted an experiment leading to experimental support S in a situation in which I had no prior preferences. (p. 36).

Then it follows that posterior support = prior support + experimental support, and we are on our way.

With this concept, and with the orientation to modification of support in light of new data, the theory acquires a strong Bayesian flavor. If its major claim to

superiority over Bayesian methods is its restriction to aleatory probabilities and the rejection of epistemic probabilities, the chief advantage of likelihood theory over the Neyman-Pearson approach is thus its orientation to the support offered to different hypotheses by the data. It would consequently appear more relevant in theoretical research settings, whereas Neyman-Pearson theory has most relevance for applied problems of testing in repetitive contexts. Edwards in fact summarizes his theory as “Bayesian inference without the priors,” “Neyman-Pearson inference without the errors” (p. 176). Interestingly, likelihood theory has not found a very enthusiastic reception, perhaps in part because those who were dissatisfied with Neyman-Pearson theory have been more attracted by the possibilities for Bayesian theory as a model of human inference.

10.5 Shafer's Theory of Belief Functions

In the 1960s, Arthur Dempster (1967, 1968) published a series of papers on upper and lower probabilities, which can be thought of as upper and lower bounds for probabilities whose values cannot be specified exactly, in a tentative exploration of their utility in statistical inference. Subsequently, Glenn Shafer picked up these concepts and developed them into an impressive theory of belief functions. Among the reasons for giving Shafer's theory rather detailed consideration here is that it proves to be a generalization of other theories; Bayesian theory, in particular, drops out as a special, rather bizarre, case. The theory of belief functions is thus interesting not only in its own right but also for the light it throws on its rivals. Like Cohen's (1977) theory of inductive, or Baconian, probability, and Shackle's (1961) theory of disbelief, or surprise, Shafer's theory is radical, in incorporating nonadditive probabilities.

Shafer begins by defining a *belief function* for some universe of possibilities, Θ , which he calls a *frame of discernment*. Belief functions can be construed as one-sided betting rates (to facilitate the comparison with Bayesian probabilities), and they commonly arise from testimony.

Suppose, for example, that Betty tells me a tree limb fell on my car. My subjective probability that Betty is reliable is 90%; my subjective probability that she is unreliable is 10%. Since they are probabilities, these numbers add to 100%. But Betty's statement, which must be true if she is reliable, is not necessarily false if she is unreliable. From her testimony alone, I can justify a 90% degree of belief that a limb fell on my car, but only a 0% (not 10%) degree of belief that no limb fell on my car. (This 0% does not mean that I am sure that no limb fell on my car, as a 0% probability would; it merely means that Betty's testimony gives me no reason to believe that no limb fell on my car.) The 90% and the 0%, which do not add to 100%, together constitute a *belief function*. (Shafer, 1990, p. 474)

Although additivity is not included in the defining properties of a belief function, the other two usual axioms are retained: Zero belief is allocated to the assertion of the null set, and the total belief over Θ is normalized to 1. The simplest illustration is the *vacuous* belief function. Consider Shafer's example, the question of whether

there are living beings in orbit around the star Sirius. If we have no evidence bearing on the question one way or another, we will accord zero belief to each possibility, though still a belief of 1 in the whole set (yes or no). As Edwards (1972) remarks, “There is no finer way of expressing ignorance than saying nothing” (p. 59), and this is just what Shafer’s vacuous belief function does. Contrast the assignment made by a Bayesian or by the principle of indifference: In the absence of knowledge favoring one alternative over the other, we would have to assign each a probability of $\frac{1}{2}$. Yet this is just the sort of example that leads easily to contradictions; for if we introduce another element into the partition—for example, the question whether there are even planets around Sirius—then we would have to make a different probability assignment, perhaps allotting a probability of $\frac{1}{2}$ to the possibility of planets but no life, which would seem to imply that, if Sirius had a solar system, it would be guaranteed to have life.

By progressively imposing restrictions on belief functions, we can construct a hierarchy of support functions. Belief functions themselves are the most general; at the opposite end are *simple support functions*. Roughly speaking, a simple support function represents a situation where all the evidence points precisely and unambiguously to a single alternative. *Separable support functions* are obtained when simple support functions are suitably combined—or, put the other way around, when the evidence can be decomposed into components that are homogeneous with respect to the frame of discernment. *Support functions* are then obtained by “coarsening” the frame of discernment for separable support functions, neglecting some of the distinctions they contain. Shafer believes that the class of support functions is sufficient to represent the impact of any body of evidence on a frame of discernment. The residual class of belief functions, however, called *quasi-support functions*, turns out to be especially interesting. But to appreciate their role in the theory, and its relation to classical and Bayesian probability, a few more details are needed.

First, the rule for combining evidence: Shafer adopts a procedure proposed by Dempster, although he traces it originally to Lambert’s *Neues Organon*, published in 1764 (See Chap. 3.). Dempster’s *orthogonal sum* is essentially a multiplicative function of support, normalized to correct for sets to which no belief is assigned. Thus consider a case of conflicting evidence. We have two simple support functions, S_1 and S_2 , in a particular frame of discernment. Let A be the set of all points assigned positive probability by the first, and B the corresponding set for the second; and suppose $A \cap B = \emptyset$. If the support given to A by S_1 is s_1 and that given to B by S_2 is s_2 , then we may construct the following table.

$1 - s_2$	Committed to A	Uncommitted
s_2	Committed to \emptyset	Committed to B
	s_1	$1 - s_1$

Now since the amount s_1s_2 is committed to the null set— S_1 and S_2 commit no belief to common points—the rest of the amounts may be normalized by multiplying by the factor $1/(1 - s_1s_2)$; the quantity s_1s_2 measures the *weight of conflict*. Hence,

$s_1(1 - s_2)/(1 - s_1s_2)$ will be committed to A alone, $s_2(1 - s_1s_2)$ to B. (Notice that being certain of both A and B is excluded.) The incompatibility of the evidence is reflected in the fact that with the introduction of the second body of evidence, the support for A is reduced by the factor of $(1 - s_2)/(1 - s_1s_2)$.

Next, we need the concept of weight of evidence. Shafer wants this quantity to be additive and to take any nonnegative value; these criteria together dictate a logarithmic function. Shafer adopts the simplest: $w = -\ln(1 - s)$ is the *weight of evidence* needed to produce the degree of support s .

Now, substituting $1 - e^{-w(A)}$ for s_1 and $1 - e^{-w(B)}$ for s_2 , we have, after simplification, the following support for A and B:

$$S(A) = \frac{e^{w(A)} - 1}{e^{w(A)} + e^{w(B)} - 1},$$

$$S(B) = \frac{e^{w(B)} - 1}{e^{w(A)} + e^{w(B)} - 1},$$

If either $w(A)$ or $w(B)$ is infinite, the support for the other set will be zero. But suppose we let both $w(A)$ and $w(B)$ approach infinity simultaneously while preserving a constant difference Δ . Then the support for A will tend to $1/(1 + e^\Delta)$ and S(B) will tend to $e^\Delta/(1 + e^\Delta)$. The interesting result is that these degrees of support add to 1, but they do not constitute a support function. Shafer proves, in fact, that any additive allocation of support—including, in particular, the chances of classical theory—constitutes a quasi-support function, which represents this paradoxical situation of infinite contradictory weights of evidence.¹⁰ Shafer offers the following observation by way of clarifying the surprising relegation of chances to the peripheral domain of quasi-support functions: the evaluation of chances requires infinite evidence, namely observation of relative frequencies in an infinite series of independent trials.

One could ask for no better example of infinite, precisely balanced, and unobtainable evidence.

Chances, then, are essentially hypothetical rather than empirical and can seldom be translated directly into degrees of support. (Shafer, 1976, p. 202)

¹⁰Shafer also defines the plausibility of a proposition as

$$Pl(A) = 1 - S(\bar{A}).$$

The difference between the plausibility of a proposition and its support

$$Pl(A) - S(A) = 1 - S(\bar{A}) - S(A) = 1 - [S(\bar{A}) + S(A)]$$

measures the degree to which both A and \bar{A} are compatible with available evidence. Pearl (1988) criticizes Shafer's theory here on the ground that, in any situation where $P(A) + P(\bar{A}) = 1$, the interval $Pl(A) - S(A)$ collapses to zero, and no future conflicting evidence can ever widen it to reflect the conflict. However, as we have just seen, the situation where $P(A) + P(\bar{A}) = 1$ is one where we already have infinite contradictory weights of evidence, so it is not clear how any further conflicting evidence could arise.

Chances, of course, still play an important part in plausible reasoning, but in Shafer's theory—as in the others—they are not used as evidence; rather they form part of the model; they are simply assumed. Statistical problems are distinguished both by the peculiar simplicity and homogeneity of their evidence—frequency counts or collections of homogeneous measurements—and in their high degree of specificity. Regarding the latter characteristic, Shafer remarks:

Unlike a scientific theory, a statistical specification always has a very narrow range of application. Hence, it can never have the kind of general and surprising success that can vindicate a scientific theory. And unless it itself derives from some broader theory or knowledge, its inability to derive positive support from the process it is alleged to govern means that it must always remain quite provisional. It is from this essentially provisional nature of most statistical specifications that we derive our general distrust of arguments based on statistical models. (p. 279)

A theory of uncertain inference involves procedures both for assessment of probabilities of individual propositions and for the combination of different bodies of evidence. Though Shafer gave little attention to the former problem in his first book, he has been addressing it in more recent work (Shafer, 1982; Shafer & Tversky, 1985) by constructing canonical examples as standards for evidential assessment. For these, he has proposed the device of randomly coded messages. The reliability of a witness, for example, can be modeled by a machine that has two modes of operation, reliable and unreliable; it has probability p of being in the former mode, where it produces only true messages, and probability $1 - p$ of being in the latter, where the meaningfulness of its messages is completely unpredictable.

Krantz (1982) cautions that Shafer's procedure yields only ordinal, and not really numerical, measures, as it might appear. He offers an analogy which would be apt for much psychological measurement as well: “Suppose we attempted to measure the ‘degree of aesthetic pleasure’ from viewing a painting by comparing the experience to the sweetness of a graded series of sucrose solutions of known concentration” (p. 347). If this aspect of Shafer's theory is the least satisfying, it is just as well, since it is the least original and central. Though the construction of canonical examples is a worthwhile endeavor—as it would be also for likelihood ratios—I am for the time being more comfortable with the modest statement in his book (1976):

I do not pretend that there exists an objective relation between given evidence and a given proposition that determines a precise numerical degree of support. Nor do I pretend that an actual human being's state of mind with respect to a proposition can ever be described by a precise real number called his degree of belief, nor even that it can ever determine such a number. Rather, I merely suppose that an individual can make a judgment. Having surveyed the sometimes vague and sometimes confused perception and understanding that constitutes a given body of evidence, he can announce a number that represents the degree to which he judges that evidence to support a given proposition and, hence, the degree of belief he wishes to accord the proposition. (p. 20)

With respect to the combination of evidence, Shafer's theory offers two choices. We may form support functions for the individual observations and then combine them, in terms of epistemic probability, by means of Dempster's rule, or we may combine the observations using the product densities for independent trials, thus

obtaining, by aleatory probability, a single compound trial for which a support function may be derived. The two results will not generally be the same, owing to the fact that epistemic combination allows us to preserve dissonance, whereas aleatory combination forces the final support function to be consonant.

Bayesian belief functions constitute a subset within Shafer's theory; they can be interpreted as a special case where an unreliable witness always lies (Shafer, 1990). They are constrained to satisfy additivity so as to behave like chances, and hence fall in the category of quasi-support functions. Shafer agrees that when chances are *known*, then it is reasonable to adopt them as one's degrees of belief, but more often chances are stipulated rather than known, and are included in the frame of discernment rather than the belief function.

Just as Bayesian distributions are a special case of Shafer's belief functions, so Bayes' Theorem is a special case of his rule for combining evidence. It is, in particular, the appropriate formula to use when our belief function is actually Bayesian. But it contains special features which are not always desirable. It requires us, first, to stretch what is often merely a few "crumbs of information" into a detailed and precise prior distribution, whereas Shafer's procedure does not require us to extend our commitments of belief beyond what we have specific evidence for. Second, Bayesian theory treats old and new evidence asymmetrically; the prior distribution is a distribution of opinion, of partial belief, but the new evidence is always a fact, assumed to be given with certainty. The asymmetry does not exist in Shafer's theory. The latter, being more general, can use Bayes' Theorem; or it can combine new evidence, expressed by a non-Bayesian belief function; or it can simply describe the support represented by the new observations themselves.

It follows from the fact that Bayesian theory treats all beliefs as chances that it is insensitive to the distinction, made in Shafer's theory, between aleatory and epistemic combination of evidence. Shafer's theory would, like Bayesian theory, assign absolute degrees of belief (within a given frame of discernment) to hypotheses; but in both the allocation of probabilities and the rules for their combination, it would allow much more flexibility than the Bayesian model. Despite the fact that the theory could be used as a decision theory, it is not constrained to a decision orientation in epistemic contexts, and could therefore clearly avoid the problems associated with a behavioral orientation and reliance on error probabilities.

Both estimation and hypothesis testing can be accommodated in Shafer's theory. He construes estimation, not as "the narrow and by itself pointless task of using evidence to choose a particular $\theta \in \Theta$ as one's guess or 'estimate' for the true value of θ ," but as "the problem of using observations to assess degrees of support for subsets of the set Θ " (1976a, p. 262). The various aspects of the statistical specification are included in the frame of discernment: which variables to include in the problem, which to assume as random, what density to assume to govern their behavior, etc. These are all matters of supposition rather than claims for which we ordinarily have direct empirical evidence. Shafer believes that his method of support is addressed to the same kinds of questions which Neyman-Pearson confidence intervals are used in practice, somewhat less satisfactorily, to answer.

Tests of hypotheses arise as a special case of deciding when to enlarge our frame of discernment. The preceding paragraph indicated how statistical specifications tend in general to involve more or less extravagant assumptions unsupported by evidence; the same is true, in varying degrees, of any other models or frames of discernment. What determines how loose or how tight a frame we choose? Two general observations may be made. (a) A larger, more comprehensive frame is not always to be preferred to a smaller one, for it is always possible to enlarge the frame sufficiently to reduce our evidence to “a collection of nullities” (1976a, p. 276). In particular, it is usually possible to enlarge a statistical specification to include alternatives which are very close to the hypothesis in question, so that differential support is rendered nugatory. As Shafer says, we want our frame to be tight enough that our evidence will “interact in an interesting way” (p. 280). (b) But on the other hand, as we tighten the reference frame, we increase the internal conflict; an excess of conflict is just what prompts a search for a larger frame.

These considerations shape Shafer’s procedure for significance testing. We form the expression for the weight of conflict contained in our observations under some statistical specification; if the degree of conflict is “too large,” we reject (some aspect of) the specification, in favor of a wider frame. In general, the weight of conflict will increase without bound as the number of observations increases; hence, some rejection criterion is needed. To this end, we may follow the traditional practice, comparing it with the weight of conflict to be expected with a given number of observations under the model, calculated from the ordinary product of chance densities; if the chance of some large amount of conflict is less than some prescribed value, we can reject. This procedure, of course, copies the standard significance test, with the exception that the test statistic, instead of being selected on the grounds of mathematical convenience, is determined in a logical way, and represents the same feature of any body of data.

Shafer wants to allow the possibility of indefinite refinement of the frame of discernment, except that it must stop short of formalizing all knowledge, and especially the evidence that bears on it. It is implicit in the Bayesian approach, on the other hand, that new evidence is automatically assimilated to the frame of discernment. This characteristic, he argues, is what puts them in the position of having to assign beliefs on the basis of no evidence: At some point, there will be no prior evidence to justify the required allocation of belief. Consequently, he insists on a distinct logical status, as well as a restricted scope, for the frame of discernment.

The question can also be considered about the relation between Shafer’s theory of belief functions and L. J. Cohen’s theory of inductive probability. At the time of writing *A Mathematical Theory of Evidence*, Shafer had available only an earlier version of Cohen’s theory. Cohen’s rules for inductive probability appear to coincide with Shafer’s support functions, though Shafer (1976) criticizes Cohen for not having allowed the possibility of dissonant evidence. In his subsequent work, *The Probable and the Provable*, Cohen (1977) puts forth as a special virtue of his theory that it can accommodate contradictory evidence. The point is a crucial one in scientific contexts; since, as Feyerabend (1975) observes, no theory of anything has ever been fully adequate, any theory always conflicts with some existing piece of

evidence. Traditional, additive probability would thus hold the probability of any theory, on the whole body of available evidence, to be 0. If probability is taken as a measure of support, we need to allow a positive probability on data that may include a contradiction. Cohen claims his inductive probability will do just that. But it is not quite clear that his theory will actually allow the combination of dissonant evidence. He appears to say instead that the emergence of a contradiction tells us that we have used the wrong criteria of evidence in at least one case. “The emergence of apparent support for a contradiction forces us to regard one . . . method of support-assessment as requiring readjustment” (1977, p. 179). And: “[The emergence of contradictions] enables us to elucidate how the actual progress of science in a particular field of inquiry imposes a continuing local readjustment in our criteria of evidential support” (p. 181). Though I may be interpreting “evidential criteria” too broadly, it would appear unnecessarily drastic to impugn them automatically in the event of a contradiction. In general, however, Cohen’s theory of inductive probability would appear to fit in as a restricted case of Shafer’s theory of belief functions.

10.6 Recent Work on Reasoning in Philosophy and Artificial Intelligence

Though the work on nonadditive probability by Shackle (1961) and L. J. Cohen (1977) has attracted little notice, Shafer’s work has been picked up with gratifying speed and energy by the artificial intelligence community. Shafer (1990) attributes the recognition to the similarity of his system to the less elaborated approach used by Shortliffe and Buchanan (1975/1990) for their automated medical reasoning system MYCIN, but he may be modest about the role of his own networking in this country, which far exceeds anything done by the two British theorists. Shackle has the additional disadvantages of having offered a theory that was less formal and axiomatized, and hence less appealing to mathematicians and philosophers, and of simply having been out of his time: At that point, all the initial simple-minded projects of artificial intelligence were still widely regarded as full of promise.

The response of the Bayesians, who were most directly threatened by Shafer’s theory, has led to a remarkable dialogue. A result of particular note is the collaboration of Shafer and Judea Pearl, perhaps the leading Bayesian theorist today, in editing their landmark collection, *Readings in Uncertain Inference* (1990). The friendly rivalry exhibited therein is wholly out of keeping, it might be noted, with the previous acrimonious history of probability and statistics. Specifics of the debate will require digressing briefly, however, into some concepts from current work in artificial intelligence.

10.6.1 Some Recent Concepts from Artificial Intelligence

One of the principal conclusions of Howard Gardner's *The Mind's New Science* (1985) is what he calls the computational paradox: The further the computer model of the mind is pushed, the less the mind looks like a computer. As many observers (e.g., Dreyfus, 1979) have noted by now, early work in artificial intelligence tended to focus on tasks that were difficult for people but easy for computers; more recently, attention has turned to those things, like reading or going for a walk, that come comparatively easily to humans. In response to the computational paradox, the field has moved away from the clean elegance of propositional logic and statistical inference to the messy problems of semantics, of content and context, and causality. Whether the formalistic approach, which has generally been retained, is the most promising has been subject to debate (cf., e.g., Reiter, 1967/1990). Several related developments can be mentioned by way of illustration.

The concepts of everyday discourse differ from those of mathematics and logic, both intensionally and extensionally, with respect to their heavy dependence on context. Winograd and Flores (1986) illustrate the intensional ambiguity in even a concrete noun such as *water*. To the question, "Is there any water in the refrigerator?" the reply "Yes, in the cells of the eggplant" might be appropriate if the context of the inquiry were moisture damage to photographic plates stored there. Similarly, Medin and Thau (1992) observe, as we saw in Chap. 7, that, despite the apparent logical similarity of the utterances, we mean something drastically different in saying "A butcher is like a surgeon" and "A surgeon is like a butcher."

Extensionally, the concepts of everyday language are also not well-defined sets, all of whose members are interchangeable with respect to the attributes defining the sets. In the psychological literature, Eleanor Rosch's work (e.g., Varela, Thompson, & Rosch, 1991) on prototypes gave (belated) theoretical recognition to the everyday observation that some instances of a class (apple, sparrow) are better exemplars than others (tomato,¹¹ penguin).¹² It is characteristic of statements about prototypes, however, that they carry unstated exceptions. And "Reasoning with exceptions,"

¹¹ Just how poor an exemplar the tomato is of fruit was made clear by the Supreme Court in 1893, when it declared the tomato to be a vegetable (*Nix v. Hedden*, 1892). The massive Tariff Act of 1883, entitled "An act to reduce internal-revenue taxation, and for other purposes"—it also prohibited the importation of contraceptives and obscene materials—imposed a 10% duty on imported "vegetables in their natural state," and other duties on specified fruits, but nonenumerated fruits were free. The defendant argued that the tomatoes he imported from the West Indies were exempt. The sole evidence submitted by either side consisted of dictionary definitions of various articles of produce, none of which the Court admitted. Mr. Justice Gray, writing for the majority, said:

Botanically speaking, tomatoes are the fruit of a vine, just as are cucumbers, squashes, beans, and peas. But in the common language of the people. . .all these are vegetables, which are. . .usually served at dinner in, with or after the soup, fish or meats which constitute the principal part of the repast, and not, like fruits generally, as dessert. (p. 307)

¹² Not that the computer need be left behind: Patricia Churchland (1991) argues that the neural nets of parallel distributed processing recreate precisely the Rosch-Wittgenstein prototypes. They "simply 'embody' the desired function as opposed to calculating it by recursive application of a set of rules listed in an externally imposed program" (p. 12).

Pearl (1988, p. 1) warns us, “is like navigating a minefield: Most steps are safe, but some can be devastating.” If anyone had occasion in everyday discourse to make the banal utterance, “Birds fly,” she or he would not be intending the statement either as the report of an exhaustive empirical survey or as a universal, essential truth, but rather as an implicit demarcation of a domain of discourse. As Judea Pearl said somewhere,¹³ “When I say ‘Birds fly,’ understand that I’m not talking about penguins, roast turkeys, baby birds, birds with broken wings, etc.”

It is obvious, in the first place, that we are heavily into communication conventions. To most investigators, these have appeared to call for different treatment from other implicative relations between concepts, but Pearl (1988) argues, with his usual flair, for the same probability-based analysis:

It is . . . hard to believe that the human mind would adopt two types of logics, one for hunting birds (statistic-based defaults) and the other for talking about birds (convention-based defaults). A decade of AI debate on whether the sentence “Birds fly” should be interpreted statistically or procedurally attests to the obscurity of this distinction. It is more reasonable to assume that reasoning patterns designed to handle communication conventions have evolved *on top* of those designed to handle empirical needs, and if this is the case, the axioms governing the latter should be adopted as canonical norms for all nonmonotonic logics. (p. 480)¹⁴

In referring to nonmonotonic logics, Pearl points to another feature of reasoning about prototypes. Classical logic is monotonic in the sense that a conclusion, once established, is not undercut by new evidence. Nonmonotonic reasoning, on the other hand, allows us to jump to potentially retractable conclusions, and in everyday discourse, which typically proceeds on the assumption that we would have been told

¹³I copied this quote sometime in the early 1990s, but was never able after that to find it in Pearl’s writings, nor did he respond to a request to help me locate it.

¹⁴If we were to distinguish pragmatic from communicative uses of language, or modes of reasoning, however, a case could be made for reversing Pearl’s hierarchy and subordinating the former to the latter. I am not sure any theory has yet taken adequate account of the social aspect of knowledge. I am not thinking just of the obvious fact that, were it not for social interaction, we wouldn’t be speaking or thinking in English (or anything else), but of the more fundamental role evidently played by other people in our own coordinate construction of a self and world. Some students of language, like Werner and Kaplan (1963), have given at least passing consideration to such phenomena as “primordial sharing” in the development of joint objects-of-contemplation (with implied communication conventions), but the experience of people like Donna Williams (1993, 1994) reminds us more radically that without connections to other people the world loses its meaning. The distinction between “my world” and “the world” attests clearly enough to the role of the social; it becomes especially dramatic in relation to the body, which inhabits both. Williams reports that until she managed to establish personal connections with other people (though she was a fluent language user), when she touched her arm, she could experience the touch objectively, as her arm being touched, or subjectively, as her hand touching something, but not both at the same time.

Autism is the extreme example, and it would be a mistake to deny its uniqueness. Yet, for all of us, our precious communication conventions are fragile; the difficulty of cross-gender communication (Tannen, 1990) is a painfully familiar reminder. A rash of breakdowns can make it a challenge for any of us to coordinate “the world” and “my world,” and threaten the web of shared meanings on which we tacitly depend. If such considerations have any implications for knowledge generally, the challenge to artificial intelligence would evidently be severe.

anything else that was relevant to the judgment at hand, nonmonotonic reasoning is clearly the rule (Reiter, 1967/1990). In subjecting plausible reasoning to a close examination, Pearl (1990b) concludes that it corresponds more closely with a logic of “almost all” than with either majority logics or support logics.

The concept of causality is also implicated in the analysis of everyday reasoning. Shoham (1990) sees causal reasoning in general as nonmonotonic. In his example, if we say that turning the ignition causes the car to start, we are assuming that the battery is not disconnected, that there is not a banana stuck in the tailpipe, and so on through an endless list of potential interferences which it would be infeasible to check. Interestingly, he contrasts the reasoning of physicists, who he says don’t use the concept of causality because they simply assume, or stipulate, all the necessary conditions to be known, but that seems to be just what the concept of causality accomplishes for us, on Shoham’s account, in everyday discourse.

Pearl (1988), who has undertaken a thoroughgoing Bayesian analysis of reasoning, offers a provocative analysis of causality in terms of its organizing knowledge modularly for the sake of efficiency.

If I ask n persons in the street what time it is, the answers will undoubtedly be very similar. Yet instead of suggesting that the persons surveyed or the answers evoked somehow influenced each other, we postulate both the existence of an invisible central cause, the correct time, and the commitment of each person to adhere to that standard. If I wish to predict the response of the $(n + 1)$ -th person I do not need to go back and consult each of the previous n responses; it is enough to consult the previously computed estimate of the actual time, then guess the next person’s response while accounting for possible inaccuracies. Conversely, after hearing the $(n + 1)$ -th answer, if I need to identify the individual owning the most accurate watch, it is sufficient for me to update the estimate of the correct time, then find the person whose answer lies closest to that estimate. Thus, instead of being a complex n -ary relation among the individuals involved, the causal model in this example consists of a network of n relations, all connected in a starlike pattern to one central variable (the correct time), which serves to dispatch information to and from the connecting variables. Psychologically, this modular architecture is much more pleasing than one that entails communication between variables. Since each variable is affected by only one source of information (i.e., the central cause), no conflict arises; any assignment of belief that is consistent with the central source will also be consistent with the beliefs assigned to other variables, and a change in any of the variables can communicate its impact to all other variables in just two steps.

Computationally speaking, such invisible causes are merely names given to storage places, which, by holding partial results, facilitate efficient manipulation of the visible variables in the system. They encode a summary of the interactions among the visible variables and, once calculated, permit us to treat the visible variables as though they were mutually independent...

In fact, this sort of independence is causality’s most universal and distinctive characteristic. In medical diagnosis, for example, a group of co-occurring symptoms often become independent of each other once we identify the disease that causes them. When some of the symptoms directly influence each other, the medical profession invents a name for that interaction (e.g., *syndrome*, *complication*, or *clinical state*) and treats it as a new auxiliary variable, which again assumes the modularization role characteristic of causal agents—knowing the state of the auxiliary variable renders the interacting symptoms independent of

each other. In other words, the auxiliary variable constitutes a sufficient summary for determining the likely development of each individual symptom in the group; additional knowledge regarding the states of the other symptoms becomes superfluous. (pp. 383–385)

In fact, the concept of causal modularization supports the feasibility not only of reasoning, but of Pearl's Bayesian theory of reasoning. Dempster and Kong (1988/1990) argue that causality resists probabilistic modeling because it operates in narrowly specified contexts, where conditions are held constant; probabilities, in contrast, average over those conditions. But Pearl's modular concept of causality would evidently allow him to reply that these modules function precisely to isolate such contexts, within which probabilities can then be applied.

10.6.2 Bayesian versus Dempster-Shafer Formalisms

We have already seen what is perhaps the fundamental distinction between the Shafer and Bayesian approaches, having to do with the expression of ignorance. Shafer (and many others, as we have seen) criticize Bayesians for assigning probabilities arbitrarily in the absence of any information. Indeed, he denies that the probabilities sought by Bayesians necessarily exist:

In order to define the probability that a person has a disease, we must either specify a population or else specify other evidence and draw an analogy between that evidence and the situation where we draw a person at random from a population. If we succeed in this, then we may speak of the probability that the person has the disease; if we do not succeed, there simply is no probability. (Shafer, 1990, p. 481)

In denying the existence of probabilities, Shafer is obviously speaking of them in an objective sense, which carries no weight for a Bayesian. Pearl (1990a) is quite willing to supply any reasonable values for missing parameters so he can get on with the computations, and he, in turn, criticizes Shafer's theory for “compromising its inferences” (1988, p. 457) by resting content with incomplete specifications.

If Shafer's theory has the advantage in handling total ignorance, Pearl (1988, 1990a) believes it fails in handling certain kinds of partial knowledge, in particular those exception-plagued prototypes. Suppose we are given the following three propositions:

- r_1 : Penguins normally don't fly;
- r_2 : Birds normally fly;
- r_3 : Penguins are birds;

with respective degrees of belief $m_1 = 1 - \varepsilon_1$, $m_2 = 1 - \varepsilon_2$, $m_3 = 1$ (ε_1 and ε_2 small), and we want to assess the support for the proposition that Tweety can fly, given that she is a penguin and a bird. On the basis of r_1 alone, $S(\text{Fly}) = \varepsilon_1$. If we add r_2 , support for the proposition that she can fly, on the basis of r_1 and r_2 , is measured by $\varepsilon_1(1 - \varepsilon_2)$; the quantity $m_1m_2 = (1 - \varepsilon_1)(1 - \varepsilon_2)$ measures the conflict of evidence between the two rules; hence normalization gives

$$S(\text{Fly}) = \frac{\varepsilon_1(1-\varepsilon_2)}{1-(1-\varepsilon_1)(1-\varepsilon_2)} = \frac{\varepsilon_1 - \varepsilon_1\varepsilon_2}{\varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2} \approx \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}.$$

This ratio of small quantities can itself be quite large, with the disturbing implication that adding the information that Tweety is a bird and that birds normally fly has the effect of substantially increasing support for the proposition that Tweety the penguin can fly. It appears as though normalization forces monotonicity.

Shafer's (1990) reply points to the hidden dependencies between the various probabilities in the problem. If we think of Tweety, as Pearl implicitly does, as having been randomly sampled from some population of birds, then the uncertainty involved in that random selection is common to the uncertainty in both the proportion of birds in the population that fly and to the proportion that are penguins. Appropriate application of belief theory would require redefinition of the frame of discernment. More generally, Shafer sees his approach as inappropriate for situations where the evidence consists of fragmentary information about a single probability distribution, and most appropriate where it consists of specific items. "If there is sufficient evidence on which to base detailed probability judgments in analogy to frequency knowledge, the Bayesian approach will be far more useful" (Shafer, 1990, p. 479).

Pearl (1988, 1990a) further charges Shafer's theory with difficulty in handling the transitivity of inference. He offers the following two rules:

r_1' : If the ground is wet, then it rained last night.

r_2' : If the sprinkler was on, then the ground is wet.

The obvious problem is that if we observe that the sprinkler was on, we can immediately infer that it rained last night. Pearl implies that Bayesianism circumvents the difficulty by virtue of being a model-based, or semantic, rather than a rule-based, or syntactic, system: A properly constructed model of the world will include statements acknowledging the sprinkler as an alternative explanation for wet ground. In denying Shafer's theory that out, he implicitly treats it as rule-based, as a purely syntactic formalism, which always combines evidence according to the same fixed rules.

I see no reason to suppose that Shafer would endorse this characterization of the difference between their approaches. In fact, I see no distinction myself in these terms. Bayesian theory is indeed more restricted and rigid in the formulas it can apply, and Shafer depends no less than Pearl on his knowledge of the world in constructing and working with belief functions.

What, finally, of Shafer's claim to have subsumed the Bayesian formalism as a special case? How, in fact, could there be any debate between them? "The answer," according to Pearl (1990a),

is that while every additive probability function is indeed a special kind of a belief function, the Bayesian analysis concerns not one but a family of such functions, and not every family of probability functions can be represented by a single belief function. For example, if we interpret IF-THEN rules as statements of conditional probabilities, the family of probability

functions satisfying these statements normally cannot be represented by a single belief function. Moreover, even if each rule in isolation can be represented as a belief function, combining these by Dempster's rule yields a belief function that bears little correspondence to the family of probabilities restricted by the conjunction of the rules. (p. 570)

Pearl is presumably thinking here of cases like the Tweety and sprinkler examples, to which Shafer has already objected to Pearl's use of his theory.

Wherever the debate may go from here, we may notice several interesting features of its present shape. In the first place, Bayesianism obviously isn't what it used to be. The changes have come about as workers, especially in artificial intelligence, have moved away from the simple models and toy problems of philosophers, and of the cognitive psychologists who followed their lead, toward accommodation to actual human reasoning. As important philosophically as is the issue of the representation of ignorance, a more serious practical obstacle in application of Bayesian theory to real-world problems has been the necessity for complete specification. It is hardly normative, Shafer and Srivastava (1990) suggest, for an individual to have so many well-defined preferences as are required by Savage's axioms—for instance, an ordering of all possible types of report that could be issued by an accounting firm or a ranking of all members of one gender or the other as possible partners. Pearl (1986/1990) articulates the problem clearly himself:

If we need to deal with n propositions, then to store $P(x_1, \dots, x_n)$ explicitly would require a table with 2^n entries—an unthinkably large number, by any standard... Human performance, by contrast, exhibits a different complexity ordering: probabilistic judgments on a small number of propositions (especially two-place conditional statements such as the likelihood that a patient suffering from a given disease will develop a certain type of complication) are issued swiftly and reliably, while judging the likelihood of a conjunction of many propositions entails a great degree of difficulty and hesitancy. This suggests that the elementary building blocks which make up human knowledge are not the entries of a joint-distribution table but, rather, the low-order marginal and conditional probabilities defined over small clusters of propositions.

Further light on the structure of probabilistic knowledge can be shed by observing how people handle the notion of independence. Whereas a person may show reluctance to giving a numerical estimate for a conditional probability $P(x_i|x_j)$, that person can usually state with ease whether x_i and x_j are dependent or independent, namely, whether or not knowing the truth of x_j will alter the belief in x_i ...

This suggests that the notions of dependence and conditional dependence are more basic to human reasoning than are the numerical values attached to probability judgments. (pp. 367–368).

Consequently, Bayesian theory, in Pearl's hands, is now concerned with graphs, trees, and networks exhibiting the structure of inference, and Bayes' Theorem itself makes hardly an appearance. Indeed, Pearl (1988, p. 15) approvingly quotes Shafer that “probability is not really about numbers; it is about the structure of reasoning.” According to Pearl (1988), the days when we would “assign to each proposition a numerical measure of uncertainty and then combine these measures according to uniform syntactic principles” (p. 2) are past.

Incontestably Pearl has made important contributions to the study of reasoning under uncertainty. His reformulation of causality alone is a significant achievement. But what has happened to statistical inference? Pearl reveals his true Bayesian colors when he disagrees with Reiter (1967/[1990](#)) or Shafer ([1990](#)) that prototypical reasoning cannot be modeled statistically, but statistics rarely appear explicitly in his own work. In the modeling of inference, they have become as peripheral as they are in Shafer's theory. Hence I would take his recent work as confirmation of the fundamental thesis of this book: that inference is not statistical, and there is no reason to say that it should be.

There is obviously still room for Pearl to disagree, by emphasizing that frequency distributions really do tacitly underlie all his (our) probability assessments, and to counter that I am merely flogging a dead horse, in attacking an earlier, more simple-minded version of Bayesianism. Before a call goes out to the SPCDH,¹⁵ however, I want to say some more about the formalization of inference and about formalization more generally.

A second feature of the debate between Shafer and Pearl resides in what the two theories have in common. The rather extensive overlap between them, especially in view of the radicalness of Shafer's theory, suggests a correspondingly narrow space for the understanding of reasoning in the present theoretical landscape. Whether the shift in Bayesian analysis came in response to Shafer's fundamental challenge or simply to the ineluctable requirements of artificial intelligence, Shafer's approach of downplaying the importance of numbers and focusing on the structure of inference has been widely adopted. More fundamentally, Shafer's work participates in the project which the Bayesians initiated, being an attempt at a more adequate formalization of human reasoning. Both of them thus share a commitment to the value of that task.

The impressive advances of just the last two or three decades make this an awkward time to raise questions about the potential of AI, especially when such questions can easily be misunderstood. I don't hold to any romantic notions about the sanctity or inscrutability of human reason; it is obvious, furthermore, that we stand to learn a lot from such efforts, if only indirectly through Gardner's "computational paradox." On the other hand, a fair amount of mischief has been perpetrated by appealing to formalization for purposes it can't fill—though more in, say, ethics as a formalization of conduct than in AI as a formalization of reasoning.

¹⁵ Arthur Koestler's ([1964](#)) experience in criticizing behaviorism was very much like Blanshard's ([1962](#)) in criticizing logical positivism. Pursuing the verifiability theory of meaning, like a mechanical hare, through seven successive stages of modification, Blanshard was told at each step that the version he had so patiently demolished was now some years or weeks out of date, and that all the problems of the old theory had been solved. Likewise, Koestler eventually found it hard to believe that he was merely "flogging a dead horse," as the neo-behaviorists and their successors charged; he has studied the general phenomenon in an essay on the SPCDH—the Society for the Prevention of Cruelty to Dead Horses (Koestler, [1967](#)).

10.7 On Formalization

10.7.1 *Limits*

It has been a persistent dream of philosophers (Carnap is a sterling example) to achieve a full formalization of (at least some aspect of) knowledge, once and for all, so that our knowledge, or our inferences, would be rendered automatic (and therefore, presumably, automatically correct). The hope has been nourished by the example and the metaphor, of games, whose artificially restricted context gives some plausibility to the idea of complete specification. Games of chance have of course provided the inspiration and the model for the concept of statistical inference from the start, and they have posed the same problems for statistical inference as they have for artificial intelligence. It is worth noting that it is difficult to specify all the rules unambiguously in a game even as uncomplicated as horse-racing: The photo-finish camera was believed to have obviated human judgment there, until a photo was taken where

one horse's nose is seen a fraction of an inch ahead of another's, but the second horse's nose extends forward by six inches or so well ahead of that of its rival by virtue of the projection of a thick thread of saliva. (Polanyi, 1962, p. 20n)

Shafer (1988, pp. 191–192) has discussed ambiguities that arise in the understanding of rules in card games.

When we turn to application of statistical inference or other formalizations to real-life problems, the inadequacy of purely formal considerations becomes much more glaring. If there were a realm where we would expect to succeed, it would surely be physics; yet we recall from Chap. 6 that probabilistic modeling of the distribution of particles in space contains an ineradicable empirical component: It depends on the kind of particles. As we move from physics to jurisprudence or everyday life, problems of specification become more difficult in an obvious way. Piquant examples abound of the trickiness of applying ordinary logic. In addition to Pearl's (1988) sprinkler example, consider Ernest Adams' (1988) whimsical specimen of modus tollens: "If it rained, it did not rain hard; it did rain hard: therefore. . ." Best of all, Sandra Harding (1986) has demonstrated the invalidity of the classic textbook example of a syllogism in Barbara: "All men are mortal; Socrates is a man." Substitution of Cleopatra for Socrates shows that the traditional syllogism, though many generations of students failed to notice it, rests on an equivocation between two different meanings of the word *man*.

The point of such examples is not that ordinary logic is useless or invalid—it is neither; rather, that, in tacitly supplying so much in the way of background knowledge and interpretation, we tend not to realize how little formalization alone is accomplishing for us, and consequently how unrealistic our hopes are for mechanized inference. Perhaps it suffices to note our own distrust of the results of mechanized inferences in such situations. As Lempert (1988) observes, we are not reassured by the image of "jurors of the future punching into a computer their prior

probabilities and likelihood ratios after each item of evidence” (p. 64). More telling is Stuart Dreyfus’ anecdote of describing to an acquaintance how a computer program could help in making a decision about whether to buy a new car—then realizing, in response to the obvious query, that this is *not* the way he himself would approach the purchase of a car (Dreyfus & Dreyfus, 1985). In general, just as the application of any measurement procedure, at the limits of its precision, yields only random errors, so the application of any rules of inference will leave an ineluctable residue of personal judgments.

It is commonly assumed, by proponents of AI, that such difficulties in real-life applications are merely technical and contain nothing of theoretical interest. The resistance to complete formalization in physics or games of chance should give us pause; a number of critics, in fact, argue against complete formalizability in principle, for reasons that go quite beyond Gödelian paradoxes. Polanyi (1962) spoke of the tacit dimension of knowledge; Searle (1992) speaks of the Background; Winograd and Flores (1986) and Dreyfus (1992), following Heidegger, locate the reason in our embodiment. Polanyi and Dreyfus emphasize particularly the elements of connoisseurship and skill in knowledge generally. Dreyfus (1992) notes, somewhat paradoxically, that it is precisely the “higher” mental processes which are most susceptible to simulation, and those functions that we share with animals that are most difficult. Herbert Simon, he observes, included “emotions” in one of his more recent programs—but merely as *interruptors* of the ongoing, rational process. I think the most important considerations in the present context are the related circumstances that our cognition is (to varying degrees) active rather than passive and that it is goal-directed.

Avoidance of reference to concepts like goals and purposes has of course been a major design criterion in theory construction throughout psychology and AI; but Powers (1978, 1988) has shown that this longstanding teleophobia derives from a naive and misguided conception of purpose, and that, properly construed, there is no incompatibility between purpose and mechanism. Self-maintenance, as the fundamental goal of organisms (Maturana & Varela, 1980), sets the lower-order goals within the context of their organization (which includes history) and environment. The existence of goals then makes possible the concepts of meaning and relevance—and commitment. Things matter to us—they have relevance and meaning—in a way that they don’t for computers. Relevance is often defined formally in terms of shifts in conditional probabilities; but without a web of causally structured background knowledge, there is no way to get off the ground with a blind search of all possible matrices. To use Meehl’s (1954) example, there are no rules to tell us whether hallucinations of ravens is the relevant class, or hallucinations of birds, or of ravens on someone’s head. Language itself, Winograd and Flores (1986) emphasize, entails commitment; meaning involves intention. We can easily imagine a computer saying, “I take thee to be my wedded wife”; we can less easily imagine the computer *meaning* it. AI proponents may well object that with such promissory statements we are moving away from purely declarative purposes. They are right, of course, but part of the point is the way intentions are embedded in assertions posing in simple declarative garb. Shafer and Tversky (1985) interpret probability statements, in

particular, in terms of commitment, as the degree to which we're willing to stake our shirts on a proposition.

Beyond these observations, it is worth noting that the kind of discourse toward which efforts at formalization pull is a static, impersonal, universalist realism, against which recent trends in philosophy have notably been directed. Probably the most influential critic of such linguistic practices was Alfred Korzybski (1933/1958). A student of his, David Bourland (e.g., 1989), has recently created a stir with his advocacy of copula deletion. (The name of his new language, E-Prime, derives from the equation

$$E' = E - e,$$

where E is ordinary English and e represents the set of state-of-being verbs.) The elimination of the passive voice could be viewed as salutary in preventing the concealment of agency; the loss of the progressive tenses ("It is raining") seems a more needless sacrifice. More moderate voices within general semantics have protested that it is the spirit rather than the letter of the law that matters (Kenyon, 1992). Menefee (1991) proposes E-Choice as a compromise: eliminating only the usages which are objectionable in terms of reification, stereotypes, and the like. He adds that, since E-Choice is no different from ordinary English, he hopes the label will disappear with his article. All silliness aside, the use of E' tends to make explicit the implications of constructivism and thereby to frustrate formal analysis of speech.

10.7.2 *The Relation Between Philosophy and Psychology: Bayesian Theory*

The concept of commitment, in particular, is useful in revealing the relation between the logical and personalist Bayesians. In fact, the issues surrounding the general relation between the psychological and the philosophical are all represented in the distinction between these two versions of Bayesian probability theory. The connection lies in the fact that it is our acceptance of formal systems, like the axioms of logic or probability, that constitutes the link between the psychological and the philosophical—that transforms the psychological subject into the epistemic subject, to use the terms of Beth and Piaget (1961). The acceptance is a personal, psychological act, but it represents our commitment to the universal, "impersonal" validity of the norms we accept.¹⁶ The split between psychology and logic comes to seem

¹⁶Matalon (1966) seems to hold that the important question is whether the rules of probability theory enjoy *general* acceptance, even as he acknowledges the difficulty of answering this empirical question. Jeffreys' principles of prior probability allocation may even be too vague (and personal) to allow any sort of convincing demonstration of their general acceptance. The personalists, for their part, might face a different problem in trying to show that people accept normalized betting ratios as rational guides in situations of uncertainty or even that they should. The real question,

problematic primarily when we forget the personal nature of the act of acceptance, which has been stressed by Polanyi (1962).

It is just that split which underlies the distinction between personal and logical Bayesian probability and has cut off the logical theories, in the eyes of many critics, from useful ties to human activity. The distinction and the intelligibility we accord it reflect the idea, which is accepted by both Bayesians and their non-Bayesian critics, that knowledge is impersonal and that the personal is arbitrary. The two types of Bayesian theorists have simply come down on one side or the other of the distinction.

The logical probability of Keynes and Jeffreys represents the evidential appraisals of the “epistemic subject,” and the primary problem with it has always been that it appears to have no connection with the reasoning activity of the psychological subject. The personalists start with only the psychological subject, but Fine (1973) faults them for allowing subjectivism to enter without knowing (like the sorcerer’s apprentice) how to close the door. The metaphor might be preserved by saying that the trouble lies rather in their having opened the wrong door:¹⁷ Letting beliefs be utterly arbitrary, they are then left with the problem of having to establish or to assume that the special coherence conditions they impose are just those necessary to secure claims for epistemic relevance. The way would appear to be more promising for an acknowledgment of the personal aspects of logical probability than for elevating arbitrary beliefs to epistemic status by the imposition of certain coherence conditions. Logical probability in fact becomes personal in two places, though they are not explicitly acknowledged as such by Keynes, Jeffreys, or Jaynes. The first is in our acceptance of the principles of the theory as rational norms, and the second is in every application of them.

What is needed, in any case, is to repudiate the self-effacing equation of the objective with the impersonal and the personal with the arbitrary.

however (to amend Matalon’s evident suggestion), is not so much whether *everybody* accepts the particular rules in question as norms—as if truth were a matter of consensus, and nothing could be known until we all agreed on it—but whether *we do*, as individual aspiring knowers. The generality of acceptance may determine a community of discourse (Koch, 1976), but not essential validity, which does not depend on the results of a survey. The questions are a little abstract in any case, given that the acceptance in question is not a matter of signing one’s name to a set of principles, like a loyalty oath.

¹⁷The fortress is not an especially apt metaphor for cognition, with its implications of war on reality; but its psychological appeal is understandable in terms of the presumed relation of cognitive control to objectivity, and the ability to exclude undesirable elements from the edifice of knowledge.

10.7.3 Purposes

The introductory discussion under the heading of “Limits” may well have made it sound as though formalization were inherently otiose. Some of its functions, like the esthetic, can be disregarded in the present context, but we should consider briefly some of the epistemological and political issues involved.

Formalization is often presumed to serve the purpose of purifying cognitive or other behavior of personal, subjective elements, by providing either a guide to constructive action or an arbiter for disputes. In fact, it is adequate to neither task. Just as rules cannot be written to cover all contingencies in advance, so most disputes arise beyond the reach of the present formalization. As was noted in Chap. 1, it is a rare argument of consequence which is resolved by exposing a logical fallacy.

Given, however, the inherently impersonal character of rules, it follows naturally that one of their important practical functions, in general, is the depersonalization of authority. If decisions are to be made with unpleasant consequences for someone else, we can either diffuse the responsibility in a committee—the firing squad is the prototype—or try to make it appear as if the decision had no human origin and was merely dictated by rules which came from somewhere else. Rules, not being sufficient in themselves to determine our decisions or conclusions, require interpretation, which is a personal act. Hence, what formal systems actually accomplish is to grant an air of legitimacy post facto to inferences or other actions reached by quite different means; they serve to establish and to sanction whatever happen to be the existing informal arrangements. In the case of statistical inference, it is obvious that, while we pretend our empirical beliefs about psychology are determined by p values, in practice, if results of a significance test do not conform to our a priori expectations, we criticize our experiment rather than revise our beliefs.

The actual epistemological functions of formalization are somewhat more difficult to define, but we may take a cue from a remark by Grize (1962) on the example of contradiction:

One can wonder why, at least beyond a certain level of development, contradiction is something to be eliminated at all cost. If I simultaneously affirm p and not- p , Strawson says that, in effect, I have communicated precisely nothing. Maybe so. But, silence being golden, it is still hard to see why we should so scrupulously avoid this particular way of saying nothing. (Grize, 1962, p. 117)

On Grize’s analysis, the elimination of contradictions comprises two aspects. On the logical level, contradictions are to be avoided; our conceptions of formalization are such that we can formalize only what is free of contradictions; and the formalization then becomes our means of recognizing contradictions in natural thought. If we say that contradictions are to be resolved in thought because they are impossible in reality, it is because formalization itself surpasses the level of the subject and confers a measure of “reality” on the object. Hence, while contradictions are to be prevented in logic, they are a mechanism of cognitive development and are to be transcended in natural thought.

Formalization, if the example of contradiction is representative, thus serves at least to define a frame for continuing operations. Redner (1987), who agrees that “the ultimate sociological effect of formalization is to affirm authority structures” (p. 76), makes the point a little more cynically, noting that formalization serves to close a field of inquiry, sealing it in canonical form. Nevertheless, the process is dynamic, at least over the long run. In isolating and systematizing the essential aspects of previous achievements, formalization furnishes a milepost, marking our current state, and the next boundary potentially to be crossed. The progression is not necessarily developmental, in any cumulative sense; it may resemble more a Wagnerian “endless melody,” in the sequence of shifts from one key to another, without awareness of an overall direction or endpoint. Or it may indeed be developmental, as when it moves self-consciously in the direction of greater abstractness. The whole process may also be too slow for us to observe, in the most abstract systems. But, as all rules are subject to interpretation, interpretations are subject to change, and the rules may be adapted in turn to the evolving patterns of interpretation, lest they lose their claim of relevance to contemporary activity and suffer a radical abandonment. Levi (1949) gives detailed illustrations of the process in the field of legal reasoning.

Some formal structure is always necessary, for reflective, organized activity, and it must be held as an absolute. But as with more restricted, substantive claims, the absolutism is functional, rather than eternal. The difference is mainly that the wider the structure, the more successful is its illusion of immortality. We can conceive the possibility of alternative theories of color perception; we do not so easily entertain the possibility of alternative ontologies or scientific methodologies. With respect to the widest abstract structures, in other words, we tend to assume that the system we presently accept is the only one possible.

In the same sense that rules are necessary, however, they also happen to be unavoidable. They are contained by implication in any organized activity, cognitive or social. In mathematics, as Spencer Brown (1972) observes, we often discover that we have been following a rule for some time without having been aware of it. What is optional is that we reflect on the rules and make them explicit. Stating them helps us to see what we are doing. At the same time, it helps to make us aware of alternatives by implication. Once Euclid’s postulates were set up, it became possible to create alternative geometries by systematically violating one or another of them. And as Feyerabend (1975) observes—inverting the usual presumption of science as a rule-bound activity—the real advances come, not from following the rules, but from breaking them.

With respect to statistical inference, perhaps the best way of indicating what formalization does or does not accomplish is to pose an argument of Susanne Langer’s (1967) as a perceptual analogy. It has been known for at least 50 years that the perceived size of an object corresponds neither to the physical size of the object nor to its retinal projection. We see something in between, through what Langer calls visual interpretation. A process so comparatively simple as size perception is thus not “formalizable” by the laws of optics or any others that we know. Renaissance painters presumably lacked our sophistication; Alberti wrote as though a painting

should be the cross section of a pyramid extending from the eye to the object. “Why, then,” asks Langer, “did the great masters of the Renaissance who recommended such geometric methods not ruin their own art with them?” (p. 96). Perhaps for the same reason that Laplace didn’t use the Rule of Succession in his own scientific work: They knew better. “They played with their grids and transparencies” (p. 96), as Laplace did with his sunrise calculation, but didn’t use them, as they easily might have, as a virtually effortless, mechanical solution to the problem of two-dimensional representation. But Langer goes further: If they had, we wouldn’t call the result a work of art. What is involved in art, as in vision, is something more subtle. And I would say, as with perception, so with cognition. As paint-by-number is not art, so calculation is not thinking.

10.8 Summary of Challenges to Bayesian Theory

Before proceeding to a brief synopsis of modern neuropsychology, it will be useful to have in mind the principal problems with Bayesian theory.

1. I gave most space in Chap. 8 to the issue of *subjectivity*. The issue is more nuanced than the others mentioned here because Bayesianism can rejoin that excluding the subjective elements doesn’t make the whole process more objective.
2. Probably the oldest issue is that of *measurement*. In Bernoulli’s (1713) example, what probability of guilt is indicated by Gracchus’ having paled under questioning? It has never been clear how Bernoulli’s cases of “pure” evidence might be numerically measured. Bernoulli himself may have been hoping that de Witt’s tract on annuities might suggest a statistical basis, as Gavarret (1840) envisioned databases of medical statistics might make possible a statistical science of medicine; but not even Gavarret’s more circumscribed field of application has achieved that goal, partly because the relevant populations keep changing.
3. The most fundamental issue, I have argued, is that Bayesianism is constantly forcing us into *two-sided bets*. Recall Hacking’s (1974) example from Chap. 3: A scrap of newspaper in a railway car predicts snow tomorrow in Chicago, but the date is torn off. He assesses the probability as 90% that the paper is today’s, thus accepting a probability of 90% for snow tomorrow in Chicago, otherwise unlikely in April. But that doesn’t mean, he says, that there is a 10% chance of more moderate weather: If the date of the paper is unknown, it has no relevance for tomorrow’s weather. Bayesian theory has no way of handling such cases. This is a particular manifestation of its general difficulty in representing ignorance. In assigning a probability of $\frac{1}{2}$ to either side of a binary alternative, it makes no distinction between cases where we have a lot of evidence, evenly divided, and cases where we have none at all.
4. The rigidly incremental nature of Bayesian inference has no way to account for *nonmonotonic reasoning*. It is commonplace for exploratory research to discredit

our hypothesis in favor of a new alternative not previously envisioned. There is no “back to the drawing board” in Bayesian statistics.

5. In problems of application, the assumption of *randomness*, though it may be needed, is often not viable. (This problem is not unique to Bayesianism, but Bayesian theory is most often used for problems where randomness cannot be assumed.) A good illustration is the Monty Hall problem (https://en.wikipedia.org/wiki/Monty_Hall_problem): A contestant faces three doors, behind one of which is a car, and behind the other two is a goat. After she makes her choice, the emcee opens another door, where there is a goat, and asks if she wants to change her choice. Marilyn Vos Savant popularized the problem years ago when she argued that the contestant should change. Mathematics professors wrote irately to denounce her, but they were assuming that the emcee had opened one of the two remaining doors at random. But clearly he did not, since one third of the time he would show her the car, which would blow the game. Given the intention of showing her a goat, two thirds of the time he would have no choice, and thus, there is information in the choice he makes.

These issues taken together argue against Bayesian theory as the universally applicable approach that it is often taken to be.

10.9 Postscript on Bayesian Neuropsychology

If cognitive psychology was born Bayesian, the newer discipline of neuropsychology was Bayesian from conception. Bayesianism was the only *mathematical* theory of cognition, therefore it had to be the correct one. It is so axiomatic that little justification is thought to be needed: The “brain” is dealing with uncertainty, therefore it is Bayesian. Additionally, when neuropsychologists observe which neurons fire in response to a specific “stimulus,” they want to invert the inference to say which stimuli must have triggered an observed firing pattern; and Bayes’ theory is called on for the inversion. The experimenters assume that this is the brain’s way of checking its inferences, whereas it is but *their* way of checking inferences: They are standing outside the organism and can compare stimuli and firings. But the brain has only different neurons firing. The nervous system has no way of checking its encoding for error. Probabilistic connections or even causal correspondences are not epistemic relations (Bickhard & Terveen, 1996). Neuropsychologists have been seduced by external symbol systems into treating internal symbolization the same way: They can flip back and forth between “...” in Morse code and the letter S, but the brain has only encoding to work with. Representation, a true epistemic relation, is a long, laborious process of construction which requires action on the part of the knowing system (Werner & Kaplan, 1963), and the possibility of error (Bickhard & Terveen, 1996)—where the error criteria are not provided by the programmer.

The whole program of neuropsychology and artificial intelligence is based on a startlingly naïve epistemology: the seventeenth-century epistemology of

associationism. All knowledge is extensional; thinking is combinatorial, in the manner of Llull (Chap. 2). Memory consists of maps to extensional sets. Bickhard and Terveen quote Lenat and Guha: “Yes, all we’re doing is just pushing tokens around, but that’s all that cognition is” (1996, p. 115).

Neuropsychology and artificial intelligence commit what we could call the “age-getic fallacy”: They are utterly incapable of accounting for the origins of the symbols they are studying. Just as socialist economists declare that the problem of production has been solved—“The goods are here” (Rand, 1957, p. 1043)—and all that remains is the problem of distribution, so for neuropsychologists and cognitive psychologists, “The symbols are here.” A theory of symbols which cannot account for the origins of symbols is *literally* a nonstarter. If we wonder at the alacrity with which so idiotic a theory should have been seized upon, we get a clue from the book *Bayesian Brain* (Doya, Ishii, Pouget, & Rao, 2007): Every page is filled with mathematics of the deepest dye. Tensor analysis is not excluded from the field (Smolensky, 1990). Like statistical inference, Bayesian information processing feeds the luxuriant proliferation of mathematics—and of course, all those numbers guarantee that it must be scientific. Like the drunk looking for his keys under the lamppost, not because that’s where he dropped them, but because “it’s brighter over here,” associationist theories, based on counts, lure people, not because they hold any promise of truth, but because they offer a way to play scientist.

References

- Acree, M. C. (1979). Theories of statistical inference in psychological research: A historico-critical study (Doctoral dissertation, Clark University, 1978). *Dissertation Abstracts International*, 39, 5037B. (University Microfilms No. 7907000).
- Adams, E. W. (1988). *Modus tollens* revisited. *Analysis*, 48, 123–128.
- Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *American Journal of Psychology*, 73, 127–131.
- Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, 55, 91–107.
- Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society (Series B)*, 11, 115–149.
- Bernoulli, J. (1713). *Ars conjectandi [The art of conjecturing]*. Basel, Switzerland: Thurneysen.
- Beth, E. W., & Piaget, J. (1961). *Épistémologie mathématique et psychologie [Mathematical epistemology and psychology] Études d’Épistémologie Génétique* (Vol. 14). Paris: Presses Universitaires de France.
- Bickhard, M. H., & Terveen, L. (1996). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Amsterdam, The Netherlands: Elsevier.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57, 269–306.
- Blanshard, B. (1962). *Reason and analysis*. La Salle, IL: Open Court.
- Borel, É. (1952). *Traité du calcul des probabilités et de ses applications* [Treatise on the calculus of probabilities and its applications]. Tome 4. *Applications diverses et conclusions* (Vol. 4. Various applications and conclusions). Fascicule 3. *Valeur pratique et philosophie des probabilités* [Book 3. Practical value and philosophy of probabilities] (2nd ed.). Paris: Gauthier-Villars. (1st ed., 1939).

- Bourland, D. D., Jr. (1989). To be or not to be: E-Prime as a tool for critical thinking. *Et Cetera*, 46, 202–211.
- Chatfield, C. (1972). Discussion of Lindley and Smith (1972). *Journal of the Royal Statistical Society, Series B*, 34, 30.
- Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, 103, 105–110.
- Churchland, P. M. (1991). A deeper unity: Some Feyerabendian themes in neurocomputational form. In G. Munvar (Ed.), *Beyond reason: Essays in the philosophy of Paul Feyerabend* (pp. 1–23). Dordrecht, The Netherlands: Kluwer Academic.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1973). *Psychological probability*. Cambridge, MA: Schenkman.
- Cohen, L. J. (1977). *The probable and the provable*. Oxford, UK: Clarendon Press.
- Cohen, L. J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition*, 7, 385–407.
- Cohen, L. J. (1982). Are people programmed to commit fallacies? Further thoughts about the interpretation of experimental data on probability judgment. *Journal for the Theory of Social Behaviour*, 12, 251–274.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187–276.
- Cowles, M. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, NJ: Erlbaum.
- d'Alembert, J. le R. (1767). Doutes et questions sur le calcul des probabilités [Doubts and questions on the calculus of probabilities]. In *Mélanges de littérature, d'histoire, et de philosophie* (Vol. 5, pp. 275–304). Amsterdam, The Netherlands: Chatelain.
- Davies, C. M. (1965). Development of the probability concept in children. *Child Development*, 36, 779–788.
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, 243, 1668–1674.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38, 325–339.
- Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society (Series B)*, 30, 205–247.
- Dempster, A. P., & Kong, A. (1988). Uncertain evidence and artificial analysis. *Journal of Statistical Planning and Inference*, 20, 355–368. Reprinted in G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 522–528). San Mateo, CA: Morgan Kaufmann.
- Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.
- Dreyfus, H. L. (1979). *What computers can't do* (rev. ed.). New York, NY: Harper Colophon Books.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial intelligence*. Cambridge, MA: MIT Press.
- Dreyfus, H. L., & Dreyfus, S. E. (1985). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York, NY: Free Press.
- Edwards, A. W. F. (1972). *Likelihood*. Cambridge, UK: Cambridge University Press.
- Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York, NY: Wiley.
- Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London: New Left Books.
- Fine, T. L. (1973). *Theories of probability*. New York, NY: Academic Press.
- Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York, NY: Basic Books.

- Gavarret, J. (1840). *Principes généraux de statistique médicale, ou développement des règles qui doivent présider à son employ [General principles of medical statistics, or the development of rules which should govern their use]*. Paris: Bechet Jeune et Labé.
- Gièvre, R. N. (1976). Empirical probability, objective statistical methods, and scientific inquiry. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science (Foundations and philosophy of statistical inference)* (Vol. 1, pp. 63–101). Dordrecht, The Netherlands: Reidel.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, 2, 83–115.
- Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 513–525.
- Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, 43, 127–171.
- Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.
- Goldberg, S. M. (1966). Probability judgments by preschool children: Task conditions and performance. *Child Development*, 37, 157–167.
- Good, I. J. (1950). *Probability and the weighing of evidence*. London: Griffin.
- Goodfellow, L. D. (1938). A psychological interpretation of the results of the Zenith Radio experiments in telepathy. *Journal of Experimental Psychology*, 23, 601–632.
- Gratch, G. (1959). The development of the expectation of the nonindependence of random events in children. *Child Development*, 30, 217–227.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411–435.
- Grize, J.-B. (1962). Remarques sur les limitations des formalismes [Remarks on the limitations of formalisms]. In E. W. Beth, J.-B. Grize, R. Martin, B. Matalon, A. Naess, & J. Piaget (Eds.), *Implication, formalisation et logique naturelle [Implication, formalization and natural logic]* (pp. 103–127). Paris: Presses Universitaires de France.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge, UK: Cambridge University Press.
- Hacking, I. (1972). Likelihood. *British Journal for the Philosophy of Science*, 23, 132–137.
- Hacking, I. (1974). Combined evidence. In S. Stenlund (Ed.), *Logical theory and semantic analysis* (pp. 21–37). Dordrecht, The Netherlands: Reidel.
- Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, 101, 252–254.
- Harding, S. (1986). *The science question in feminism*. Ithaca, NY: Cornell University Press.
- Hochauer, B. (1970). Decision-making in roulette. *Acta Psychologica*, 34, 357–366.
- Hoemann, H. W., & Ross, B. M. (1971). Children's understanding of probability concepts. *Child Development*, 42, 221–236.
- Irwin, F. W. (1953). Stated expectations as functions of probability and desirability of outcomes. *Journal of Personality*, 21, 329–335.
- Johnson-Laird, P. N. (1988). *The computer and the mind: An introduction to cognitive science*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1, 134–148.
- Johnstone, D. J. (1988). Hypothesis tests and confidence intervals in the single case. *British Journal for the Philosophy of Science*, 39, 353–360.
- Joseph, H. W. B. (1916). *An introduction to logic* (2nd ed.). Oxford, UK: Clarendon Press.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.

- Keirsey, D., & Bates, M. (1984). *Please understand me: Character and temperament types*. Del Mar, CA: Prometheus Nemesis Books.
- Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, 15, 192–238.
- Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28, 107–128.
- Kenyon, R. (1992). E-Prime: The spirit and the letter. *Et Cetera*, 49, 185–188.
- Koch, S. (1976). Language communities, search cells, and the psychological studies. *Nebraska Symposium on Motivation*, 23, 477–559.
- Koestler, A. (1964). *The act of creation*. New York, NY: Macmillan.
- Koestler, A. (1967). *The ghost in the machine*. New York, NY: Macmillan.
- Korchin, S. J. (1976). *Modern clinical psychology*. New York, NY: Basic Books.
- Korzybski, A. (1958). *Science and sanity: An introduction to non-Aristotelian systems and general semantics*. Lakeville, CT: The International Non-Aristotelian Library Publishing Company. (Original work published 1933).
- Krantz, D. H. (1982). Discussion of Professor Shafer's paper. *Journal of the Royal Statistical Society (Series B)*, 44, 347–348.
- Kyburg, H. E., Jr. (1974). Propensities and probabilities. *British Journal for the Philosophy of Science*, 25, 358–375.
- Lambert, J.-H. (2002). *Nouvel organon: Phénoménologie*. [The new organon: Phenomenology] (G. Fanfalone, Trans.). Paris: Vrin. (Original work published 1764).
- Langer, S. K. (1967). *Mind: An essay on human feeling* (Vol. 1). Baltimore, MD: Johns Hopkins Press.
- Le Bonniec, G. (1970). *Étude génétique des aspects modaux du raisonnement* [Developmental study of modal aspects of reasoning]. Unpublished doctoral dissertation, École Pratique des Hautes Études, Laboratoire de Psychologie, Paris.
- Lempert, R. (1988). The new evidence scholarship: Analyzing the process of proof. In P. Tillers & E. D. Green (Eds.), *Probability and inference in the law of evidence* (pp. 61–102). Dordrecht, The Netherlands: Reidel.
- Levi, E. H. (1949). *An introduction to legal reasoning*. Chicago, IL: University of Chicago Press.
- Lonergan, B. J. F. (1970). *Insight* (3rd ed.). New York, NY: Philosophical Library.
- Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934–1945. *British Journal of Mathematical and Statistical Psychology*, 32, 151–178.
- Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. New York, NY: W. H. Freeman.
- Marks, R. W. (1951). The effect of probability, desirability, and “privilege” on the stated expectations of children. *Journal of Personality*, 19, 332–351.
- Matalon, B. (1966). Épistémologie et psychologie des probabilités [Epistemology and psychology of probabilities]. In F. Bresson & M. de Montmollin (Eds.), *Psychologie et épistémologie générales* [Psychology and genetic epistemology] (pp. 107–115). Paris: Dunod.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, The Netherlands: Reidel.
- Medin, D. L., & Thau, D. M. (1992). Theories, constraints, and cognition. In H. L. Pick Jr., P. van den Broek, & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 165–187). Washington, DC: American Psychological Association.
- Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis, MN: University of Minnesota Press.
- Mellor, D. H. (1971). *The matter of chance*. Cambridge, UK: Cambridge University Press.
- Menefee, E. (1991). E-Prime or E-Choice? *Et Cetera*, 48, 136–140.
- Messick, S. J., & Solley, C. M. (1957). Probability learning in children: Some exploratory studies. *Journal of Genetic Psychology*, 90, 23–32.
- Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, 20A, 175–240.

- Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika*, 20A, 263–294.
- Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London (Series A)*, 231, 289–337.
- Nix v. Hedden, 149 U.S. 304 (1892).
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.
- Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29, 241–288. Reprinted in G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 366–414). San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1990a). Bayesian and belief-functions formalisms for evidential reasoning: A conceptual analysis. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 540–574). San Mateo, CA: Morgan Kaufmann.
- Pearl, J. (1990b). Probabilistic semantics for nonmonotonic reasoning: A survey. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 699–710). San Mateo, CA: Morgan Kaufmann.
- Peirce, C. S. (1932). *Collected papers* (Vol. 2). Cambridge, MA: Belknap Press.
- Pepper, S. C. (1942). *World hypotheses*. Berkeley, CA: University of California Press.
- Piaget, J., & Inhelder, B. (1951). *La genèse de l'idée de hazard chez l'enfant [The development of the child's concept of chance]*. Paris: Presses Universitaires de France.
- Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy* (Corrected ed.). Chicago, IL: University of Chicago Press.
- Popper, K. R. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10, 25–42.
- Popper, K. R. (1962). The propensity interpretation of the calculus of probability, and the quantum theory. In S. Korner (Ed.), *Observation and interpretation in the philosophy of physics* (pp. 65–70). New York, NY: Dover. (Original work published 1957).
- Popper, K. R. (1968). *The logic of scientific discovery* (2nd English ed.). New York, NY: Harper Torchbooks. (Original work published 1934).
- Powers, W. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review*, 85, 417–435.
- Powers, W. (1988). *Living control systems*. Gravel Switch, KY: Control Systems Group.
- Rand, A. (1957). *Atlas shrugged*. New York: Random House.
- Redner, H. (1987). *The ends of science: An essay in scientific authority*. Boulder, CO: Westview Press.
- Reichenbach, H. (1949). *The theory of probability* (2nd ed.). Berkeley, CA: University of California Press.
- Reiter, R. (1990). Nonmonotonic reasoning. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain inference* (pp. 637–656). San Mateo, CA: Morgan Kaufmann.
- Ross, B. M. (1955). Randomization of a binary series. *American Journal of Psychology*, 68, 136–138.
- Ross, B. M. (1966). Probability concepts in deaf and hearing children. *Child Development*, 37, 917–927.
- Ross, B. M., & Levy, N. (1958). Patterned predictions of chance events by children and adults. *Psychological Reports*, 4, 87–124.
- Ryan, T. A. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26–47.
- Sarbin, T. R., Taft, R., & Bailey, D. E. (1960). *Clinical inference and cognitive theory*. New York, NY: Holt, Rinehart, Winston.
- Scholz, R. W. (1987). *Cognitive strategies in stochastic thinking*. Dordrecht, The Netherlands: Reidel.

- Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316.
- Shackle, G. L. S. (1961). *Decision, order and time in human affairs*. Cambridge, UK: Cambridge University Press.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Shafer, G. (1982). Belief functions and parametric models. *Journal of the Royal Statistical Society (Series B)*, 44, 322–339.
- Shafer, G. (1988). The construction of probability arguments. In P. Tillers & E. D. Green (Eds.), *Probability and inference in the law of evidence* (pp. 185–204). Dordrecht, The Netherlands: Reidel.
- Shafer, G. (1990). Belief functions. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 473–481). San Mateo, CA: Morgan Kaufmann.
- Shafer, G., & Pearl, J. (Eds.). (1990). *Readings in uncertain inference*. San Mateo, CA: Morgan Kaufmann.
- Shafer, G., & Srivastava, R. (1990). The Bayesian and belief-function formalisms: A general perspective for auditing. In G. Shafer & J. Pearl (Eds.), *Readings in uncertain reasoning* (pp. 482–521). San Mateo, CA: Morgan Kaufmann.
- Shafer, G., & Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Science*, 9, 309–339.
- Shoham, Y. (1990). Nonmonotonic reasoning and causation. *Cognitive Science*, 14, 213–252.
- Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23, 351–379.
- Smith, J. G. (1934). *Elementary statistics*. New York, NY: Holt.
- Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46, 159–216.
- Spencer Brown, G. (1957). *Probability and scientific inference*. London: Longmans, Green.
- Spencer Brown, G. (1972). *Laws of form*. New York, NY: Julian Press.
- Tannen, D. (1990). *You just don't understand*. New York, NY: Morrow.
- Tversky, A., & Kahneman, D. (1971). The belief in the law of small numbers. *Psychological Bulletin*, 76, 105–110.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- Wartofksy, M. W. (1971). From praxis to logos: Genetic epistemology and physics. In T. Mischel (Ed.), *Cognition and epistemology* (pp. 129–147). New York, NY: Academic Press.
- Werner, H., & Kaplan, B. (1963). *Symbol formation: An organismic-developmental approach to language and the expression of thought*. New York, NY: Wiley.
- Williams, D. (1993). *Nobody nowhere*. New York, NY: Times Books.
- Williams, D. (1994). *Somebody somewhere: Breaking free from the world of autism*. New York, NY: Times Books.
- Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex Publishing.
- Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram Media.
- Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, 18, 451–460.
- Yin, R. K. (1989). *Case study research: Designs and methods* (rev. ed.). Beverly Hills, CA: Sage.

Chapter 11

Conclusions and the Future of Psychological Research



11.1 Conclusions

11.1.1 *The Concept of Probability*

The meaning of the concept of probability has bedeviled philosophers for well over a century. The sharp split appearing in the early decades of the twentieth century between those theories interpreting probability as fundamentally a frequency and those taking it as fundamentally a degree of belief or support created from the beginning an implicit press for some sort of integration or reconciliation. The discouraging results of such efforts prompted Kendall (1949) to observe many years ago that

If some people asserted that the earth rotated from east to west and others that it rotated from west to east, there would always be a few well-meaning citizens to suggest that perhaps there was something to be said for both sides, and that maybe it did a little of one and a little of the other; or that the truth probably lay between the extremes and perhaps it did not rotate at all. (p. 115)

In recent decades the debate has “calcified,” as Shafer (1990a) puts it, “in a sterile, well-rehearsed argument” (p. 440). Frequentists can point with some satisfaction to the fact that they predominate in statistics and the experimental sciences, whereas Bayesianism has appealed predominately to those doing purely theoretical work in economics and artificial intelligence.

On the present analysis, it is understandable why any attempt to articulate or formalize the concept with any fidelity to everyday usage is bound to be frustrated: Bernoulli (1713) packaged together two concepts that really had nothing to do with each other. That assimilation by a single mathematician could not have been so successful, however, without outside help. The help was of two kinds. One was a certain felt plausibility to the assimilation: To say that the proposition “An ace will turn up when this die is cast” has probability 1/6 does indeed seem to say something

about the reliability or credibility of the statement. It could be argued that this interpretation is anachronistic, but the explicit application of probability to propositions waited, in any case, no more than a century; and I am inclined to regard Ancillon's (1794) contribution in this respect as more in the nature of a formality than a radical departure. The second boost came from the promise proffered by the assimilation: The implication that all statements that were not deductively certain—essentially all those outside of mathematics and logic, on most accounts—were, by virtue of being probable, subject to precise quantification in their uncertainty. That was the promise of statistical inference.

The assimilation was still not an easy one. In ontogenesis today, the adverb *probably* is usually acquired first, with an epistemic meaning; *probability* is typically learned in mathematical contexts; and the lexemic similarity very powerfully suggests the assimilation. Presumably, all of us went through a struggle like that of the precocious 12-year-old described in the previous chapter, though we may not remember it, and we may not have progressed any farther than he. Bernoulli himself, as we saw in Chap. 3, held back from a simple or outright assimilation; he was astute enough to notice that the applications he had in mind outside games of chance required a different calculus, and it was not clear that such probabilities could be numerically measured. Today, with the mathematical meaning of *probability* being given in everyday discourse, Bernoulli's option is hardly available developmentally.

There remains the question of how best to go about untying the knot. The most radical suggestion is that of Fine (1973), who proposes that the concept of (mathematical) probability can be dispensed with altogether. It might indeed be possible for mathematicians to get along without the word, but that move does not really solve the problem. The quantity presently named *probability* still occurs in equations, and it is appropriate to have a name for it. That would be true even if the mathematical focus were somehow to shift back to expectation as the fundamental quantity, as in the seventeenth century.

I agree with Shafer (1976) and others that nothing would be lost in referring to mathematical, additive probabilities as *chances*. Linguistic usage does not of course change just so simply with the fiat of a small group of writers. For an idea whose time has come, linguistic change in this age can be remarkably rapid, as the shift to nonsexist pronouns in the last six decades attests; but the time for this idea is still a way off, as the last section of this chapter will discuss.

I agree also with many contemporary theorists that the mathematical concept of chances can best be understood in terms of some concept of propensity. Such an interpretation accords with the way even mathematicians think about chances. As Lucas (1970) and Fetzer (1974) have suggested, we may think of propensity as the intensional definition of chances and frequency the extensional. The former gives the better—I would say the only—connection to theory, but we can preserve at least the spirit of Fine's (1973) recommendation if we allow the latter to demarcate the domain of useful application. If the extension is ill-defined (e.g., the probability in the next decade of an earthquake exceeding 8.0 or of an amendment to the

U.S. Constitution), the value of any probability calculations will be more rhetorical than scientific.

The concept of probability and its lexemic relatives can then be retained for their traditional epistemic usage. I am not especially fond of Cohen's term *inductive probability*, just because it sounds so close to the concept of statistical inference; but if *chances* were adopted for additive, mathematical probability, the extra qualifier *inductive* would be unnecessary. Probability, in its original, epistemic sense, does admit of degrees, but not of any precise numerical measurement. It is considered useful in some contexts, such as weather forecasting, to attach numbers to probability statements, and there is no harm in the practice so long as no one imagines, as some people do today, that they reflect some sort of relative frequency calculations, or in fact anything other than a particular meteorologist's informal (Cohen would say "inductive") assessment of the evidence.

These recommendations might appear simply to turn the clock back 180 years: Poisson (1837), grappling explicitly with the duality of the concept that had emerged a century earlier, used the same terms to mark the same distinction. But, having gone a lot farther in the meantime toward persuading ourselves that some monolithic concept would suffice, we now need to acknowledge instead with this pair of terms that chances and probability have virtually nothing to do with each other.

I put this solution forward with some diffidence, in view of Glenn Shafer's recent writings (1990a, 1990b, 1993, 1996) on the unity of probability. Shafer has influenced my thinking more than anyone else in this area, and I remain persuaded by his view of over 40 years ago (1976, p. 9) that the unification of chances and degrees of belief must be resisted, while he has gone on to argue precisely for unification. What I do think is sound, and important, in his recent work on this question, in any case, is his insistence that any application of probability entails the adoption of a model, or analogy, making it incumbent on users to defend their analogies, if not necessarily their ideologies. Thus, for example, if Fisher (1936) wants to use a randomization test to compare heights of Englishmen and Frenchmen, he is obliged to make a case for treating the determinants of height as though they were equivalent to a physical act of randomly assigning nationality "treatments." Or consider Good's (1990, 1992) excursion into "physical numerology": Confronted in elementary particle theory with a physical constant measured as 1.0000019 ± 0.0000044 , or -47.95 ± 0.085 , he stipulated a probability distribution for the constant in order to be able to make a statement about the probability that the constant is an integer. As a Bayesian, of course, he was speaking of a distribution of his belief about the value of the constant. Both of these are examples of what is called statistical inference, justification for which would depend in any case on justification of the probability model as an analog of a random process. And in all such cases, investigators confront the additional and quite possibly more difficult problem of showing that the assumption of a probability model contributes something useful beyond the data themselves.

11.1.2 *The Concept of Statistical Inference*

The greatest consequence of sundering chances and probability would be the dissolution of the concept of statistical inference.

The conditions for the emergence of the concept of statistical inference were complex; it required more than a dualistic concept of probability. Primarily, there had to be a perceived need, a problem for which statistical inference would be the solution. The problem was created in part by the dissolution of the concept of causality, the glue that had hitherto held things together. The metaphor of mucilage makes sense only retrospectively, of course: Before the seventeenth century the world was not a collection of fragments in need of a bonding agent. But the loss of the agential, Aristotelian concept of causality was still keenly felt. As all knowledge was assimilated to the omnivorous concept of sign, cause and effect both came to be identified with signs. There was no longer anything left to connect cause and effect but what could be observed on the surface: their frequency of co-occurrence. The aspiring knower confronted only a world of coincidental aggregates, arbitrary collections of interchangeable individuals. Our *experience* of causal connectedness in the world could then be explained only in the psychological terms of habit. Hence the skeptical problem of induction, setting the stage for Laplace's introduction of an urn full of tickets as the model for knowledge. The urn, at least in Laplace's initial formulation, was limited to a single case, that of binary characters; but it was the most important case, covering the question of presence or absence of an attribute; and quantitative variation could be added later as a mathematical nicety. The dualistic concept of probability was perfectly designed to make possible a new concept of knowledge: One referent was to the traditional conception of knowledge and its reliability, the other to random processes, such as inspecting a sample of elements from the coincidental aggregate of tickets in the urn.

It may still be wondered why, if the dualistic concept of probability caught on in the popular consciousness, the concept of statistical inference, which derived directly from it, did not. I think there are two things to be said in answer. One is that, in some sense, the concept of statistical inference did catch on, among those educated enough to encounter it in their own thought, or reading, or discussion. I refer here to the global assumption that inductive inference could be quantitatively formalized, as on an urn model, with the dualistic concept of probability. What did not "catch on" at all is actual "statistical reasoning," as a parallel type of human information processing. Here again two things can be said. One reason was surely the sheer mathematical difficulty of the calculations. The dualistic concept of probability required only the achievement of formal operations, and some mental acrobatics to make combinatorics seem applicable to epistemology. But only exceptional individuals can comfortably do Bayesian calculations in their heads in the case of a few discrete alternatives. Fewer still could handle realistic practical applications involving the continuum and integral calculus. But suppose that difficulty weren't a factor: I conjecture then that people would have experienced themselves merely as doing a certain kind of arithmetic problem in their heads, and not as engaged in a new type

of reasoning. I see no reason to believe that the latter ever had any psychological reality. That does not, of course, invalidate it. Real people don't reason syllogistically, either; at least they don't do it very well. But deductive logic *is* accepted, generally, at least as a standard of correct reasoning in such (artificial) cases. With statistical inference, on the other hand, there is no support for the assumption that it represents either an appropriate standard or a formalization of any human reasoning process.

At the time when the Bayes-Laplace theory of statistical inference was promulgated, there was no real distinction between philosophy and psychology, and this theory essentially provided a formalization of the associationism in both domains. With even Anglo-American philosophy having by now left associationism largely behind, statistical inference, as a formal psychological model, is moribund, surviving mainly in textbooks of cognitive psychology (e.g., Medin & Ross, 1992). As an informal assumption about reasoning under uncertainty, however, it will be with us for as long as we retain the dualistic concept of probability.

Fisher himself made this assumption explicitly, and late in his career:

That such a process of induction existed and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive process. (Fisher, 1955, p. 74)

It is a little hard to tell, in reviewing his life and work, to what extent he was guided from the start by the goal of formalizing induction. Though he was always sensitive to the wider philosophical implications of his work in statistics, his initial impetus seems to have been more the solution of applied problems in agricultural research, and the fascination of mathematics; the explicitly epistemological themes come more to the fore in his later work. It is, in any event, ironic that it was largely through Fisher's own work that his claim about induction was coming to be doubted by the time he made it.

The original Bayes-Laplace theory of statistical inference was no more of a logical success than the concept of probability on which it was based. Various diagnoses were made, as we noted, with the most popular pinning the blame on the principle of indifference, particularly in its application to unknown prior probabilities. A method of making statements about a population on the basis of a sample without appeal to that principle had to wait a century and a half for the genius of Fisher. Given that lapse, there is even some question whether Fisher was truly working on the same problem as Bayes. Foucault's (1973) description of the transformation of knowledge engendered by the intervening Kantian revolution would appear to make that proposition doubtful. If disciplines such as economics, biology, and linguistics had changed their form entirely as a result of that transformation, would it not be naive to treat Fisher and Bayes as intellectual contemporaries? I think the answer is negative, precisely because this “organic revolution” bypassed epistemology entirely. Early twentieth-century philosophies of knowledge were the most avowedly atomistic in history, even if Fisher himself was epistemologically more traditional than modern.

Fisher's theory was designed explicitly to conform at least to a Venn-vintage frequency theory of probability. Yet, to achieve a theory of statistical inference, he wanted, in the end, to make a traditional, epistemic interpretation of frequency probabilities, specifically to treat significance levels as a numerical measure of support for (or disconfirmation of) an hypothesis. His concept of fiducial probability was still more inscrutable, in preserving an appearance of frequency probability while not involving any frequencies at all. If such an interpretation were allowed, however, any principled objection to Bayesian theory, at least from a frequentist point of view, was lost after all, and Fisher spent all his life defending his equivocations on the interpretation of probability.¹

The concept of significance testing (and of fiducial probability) provided the means for a thoroughgoing frequentism in the work of Neyman and Pearson. But in giving up the epistemic interpretation of probabilities, Neyman soon realized that he no longer had a theory of statistical inference, but simply a statistical decision theory. From the premise, "If H_0 is true, then results R are unlikely," plus observation of R , no conclusion follows; an inference as to the truth of H_0 would not even fully qualify as plausible, as Pollard (1993) has observed. The Neyman-Pearson theory refers to a range of alternative distributions in selecting the location of the rejection region, and assumes that results far from the null value are more likely under some alternative. The Bayesians often specify a point alternative against which to compare the likelihood of the results under the null. One way or another, an extra premise must be included to warrant a plausible inference. The distinction between the two is that the Bayesian theory is concerned with the entire inference, which might then appropriately be called statistical; but under the Neyman-Pearson theory, any such inference lies outside the theory, in the realm of classical, plausible reasoning; the theory only supplies (some of) the premises. Thus, whatever inference is involved in hypothesis testing, it is not peculiarly statistical, nor is it part of the Neyman-Pearson theory. The theory pertains, as Neyman always insisted, to a decision, an overt behavioral act, rather than to inference.

The more recent attempts to formalize inductive inference by Shafer (1976) and L. J. Cohen (1977)—not just to "clean up" actual human inference, but to provide an epistemologically adequate model—have finally made it clear, even if that was not their intent, that inference and statistics have nothing more to do with each other than, say, ethics and trigonometry. For in constructing an adequate account of inference, they were led to abandon all reference to chance and random processes. Shafer's theory is abstract enough to encompass various ways of combining evidence, including Bayesian, but in that process it makes clear what an extraordinary set of conditions Bayesian inference requires.

¹It is ironic, and perhaps inevitable to retrospect, that two of the most recent attempts to grapple with the concept of statistical inference have essentially struggled, with different degrees of formality, to rehabilitate Fisher (Mayo, 2018; Weisberg, 2014). Although their approaches can boast a match with everyday usage, I cannot see that they have succeeded in overcoming the incoherence of the concepts of probability and statistical inference.

The status of Bayesian theory is a little more ambiguous: Traditional Bayesian procedures would qualify as statistical inference if anything would. Yet the feature which gives it that title is also the one which threatens its applicability: Its rule of combining evidence is the same as that for chances. Any theory of inference can of course accommodate statistical probabilities as *data*, and there is also no question about the validity or usefulness of Bayes' Theorem as a formula for combining chances. Yet as Shafer shows, constraining beliefs to behave as chances, by using Bayes' Theorem to combine beliefs, effectively allots them all infinite contradictory weights of evidence. Many cognitive psychologists and some mathematicians have assumed that the answers given by Bayesian methods were the most appropriate. Given the ambiguities of quantifying subjective probability, it might be difficult to mount a convincing counterdemonstration; and ultimately anyone is free, in any case, to declare that his or her beliefs are adequately modeled by Bayes' Theorem; but the arguments of Cohen and Shafer suggest that such circumstances will be exceptional. In sum, we might grant traditional Bayesian procedures uniquely the title of statistical inference, but essentially as an indictment: Not only do such procedures fail to model human inference, but there is no reason to say that they should. Recent Bayesian work in artificial intelligence (Chap. 10), taking seriously the task of modeling human inference, has pushed probability distributions into the background. Pearl's elegant directed acyclic graphs may still have to be cashed out in complete probability specifications at the point of application, but the extent to which statistics has disappeared from his models nicely confirms my own contention that inference is not statistical. We might well leave the last word on the subject to the person who was the first I have noticed to use the term *statistical inference* in a formal way: Keynes (1921/1973), subsequently dubbed a Bayesian, pessimistically concluded, Fisher's work still in the future, that statistics should be limited to "counting the cases," and that efforts at statistical inference should be abandoned.

Statistics, divorced from inference, remains a legitimate science, and the Neyman-Pearson theory, at least in its global outlines, remains a valid decision theory. Either may find legitimate application to any random mass phenomena—traffic on freeways or communication lines, gambling, the spread of contagious diseases, the occurrence of accidents or defects in mass production, the distribution of particles in space, or innumerable others. Certain psychological processes, such as in perception, may find random mass phenomena a useful model. But the inferences or decisions made in such situations would, or should, have nothing special about them. If any psychologists, or any other scientists, are really concerned to examine the formal structure of their inferences, they may start by consulting Shafer (1976), L. J. Cohen (1977), or Pearl (2000, 2018); I suspect few will feel the need.

11.2 The Future of Psychological Research

11.2.1 Surface Obstacles to Change

In assessing future directions for psychological research, it will be convenient to start with the more immediate and tangible obstacles to change, and to proceed to developments which lie in the near future, in either a logical or a temporal sense. Obstacles at this level are not necessarily easy to deal with, just because they are readily identified; neither are they, unfortunately, all we are up against. In a later section of the chapter, I address some of these wider (or deeper, depending on which way the paper is turned) issues.

The myth of statistical inference has achieved a power arguably to rival that of any other myth in this century, but what it has inspired is not, in my view, a beneficent creation. Its most tangible result is what no doubt amounts by now to billions of dollars' worth of research of questionable value, and the whole profession of academic psychology built around it. That is in itself the biggest obstacle to change. I don't mean necessarily anything so simple and crude as a conscious concern for job security, though that will play its part. There are also the issues of cognitive dissonance and self-esteem. If we have invested a lifetime, whether planned or completed, in traditional psychological research, it will be difficult not to resist any argument, whatever its merit, that the enterprise is fundamentally misguided.

Proposals for change are not lacking; the most idiotic of those surveyed by Gigerenzer (2018) was to replace the 0.05 criterion with 0.005. But even the sensible proposals, like improving the quality of statistics education, have a sense of futility about them like proposals for police reform. In the whole cultural and political system, every constituency (except the taxpayers, who make it all possible) has a vested interest in the status quo. As I pointed out on the first page of Chap. 1, the great virtue of the system, to those within it, is the mass scale of production it makes possible. It is not just that quality is sacrificed to quantity; as Tom Lehrer said in another context, “Quality is no object.”

The system hardly depends for its survival, however, on dissonance reduction within individuals. It is hard to underestimate, first, the role of the federal government in creating and maintaining the profession of academic psychology on the scale of the last 60 or 70 years. Federal funding of science had been negligible up to 1940; then the government hired lots of scientists during World War II. After the war Vannevar Bush—founder of Raytheon, inventor of the digital differential analyzer, and head of the wartime National Defense Research Committee—thought it was intolerable to turn them all back out on the street, and lobbied for the creation of the National Science Foundation—soon to be followed by the National Institutes of Health and all the rest. Starting with the National Defense Education Act of 1958 in response to Sputnik, the government also began pumping massive amounts of money into higher education. Such a program of massive federal support for research was new, and was thought to require some justification; court economist

Kenneth Arrow (1962),² of RAND, rose to the challenge.³ Arrow argued, along conventional lines, that since knowledge, as a public good, was nonrivalrous and non-excludable, it would therefore be underfunded by the private sector. The problem is solved, Arrow says, only “in an ideal socialist economy” (p. 617), of which he gives only the Soviet Union as an example.⁴ Nonrivalrous it may be; my knowing something doesn’t preclude your knowing it; but nonexcludability has been contested. Kealey (2021), following Polanyi (1962), argues that if indeed knowledge were nonexcludable, any discovery would be instantly available to everyone—hence Arrow’s claim that it would be underfunded. But discoveries typically occur at the frontiers of knowledge; only specialists in the field are able to grasp their significance. That grasp requires a long, costly process of acquiring the necessary skills and expertise (Polanyi’s “tacit knowledge”). As a point of fact, according to Kealey (2021), private firms average investing about 7% into basic research and development, to stay ahead of the competition. But the funding agencies have never looked back on Arrow’s paper.

A good counterexample to nonexcludability was provided by Ted Nelson in a 1988 interview. He had invented the concept of hypertext—the basis for the World Wide Web—in 1974. He said, “I was sure it would catch on like wildfire, and I was right. The only trouble was it caught on like wildfire twenty years later (Gans, 1988, p. 53).” Some knowledge producers don’t care about excludability. Kevin Carson (2008), on the copyrights page of his (self-published) *Organization Theory*, says, “Anyone found copying and distributing this book without permission will be considered a mighty good friend.” But the idea of knowledge as nonexcludable becomes downright laughable when we realize how widely knowledge is contested today, and notice how desperately the government is *trying to exclude* various knowledge claims, on the ground that they are too widely, and freely, available: that the CIA masterminded the assassination of JFK, that we are on the brink of the next ice age, that there is a link between vaccination and autism, and so on and on.

So pronounced a dominance of the market by a single source exerts its own distorting influences, but this particular source is by nature also the most conservative: It must satisfy the widest possible constituency. More importantly, the fact that government is not typically the “consumer” of the research it sponsors severs the quality control link. It is more important for granting agencies—and Congress above them—to be able to say that they are spending, say, \$20,000,000 on AIDS prevention research than to adduce worthwhile results of such research. The results will probably not be in until after the next election, anyway. Research thus tends to attract people who are good at hustling money for conventionally conceived

²I am indebted to Michel Accad for the reference.

³So did Richard Nelson, but he subsequently recanted his views, and so was not awarded the Nobel Prize with Arrow (Kealey, 2021).

⁴At the time, Paul Samuelson, in his textbook *Economics* (1964), included a graph showing that, due to its higher growth rate, the Soviet economy would overtake the U.S. economy by 1970. In subsequent editions, the intersection point was moved farther into the future, until the graph was dropped altogether (Skousen, 1997).

projects. The latest statistical techniques will enhance a proposal; more innovative methodologies will generally not.

Logically, separation of state and science would seem as obviously crucial as separation of state and religion, but that issue wasn't always obvious, either. Our high school history textbooks made it sound as though the adoption of Christianity as the state religion in the fourth century was a naturally desirable outcome—and no doubt it was perceived that way by contemporary Christians, as an advance over persecution and torture by the state (these were not the only alternatives)—but a good case could be made that it was the worst thing that ever happened to the Church, and possibly the world. It represented the triumph of a particular hierarchically oriented sect (probably the one most interested in state sponsorship) over a multitude of others; Christianity thenceforth became much more monolithic and powerful—and more sterile and corrupt. Reform from within became difficult or impossible; the explosive result, over 1000 years later, was the Protestant Reformation, and its hideously bloody aftermath (which continues to this day). Of course, one can scarcely imagine anything less frightening to modern psychologists than horrible consequences of current policies which will become manifest in 1000 years.

Science succumbed to the same temptation to state sponsorship—and it has grown predictably powerful, monolithic, sterile, and corrupt. Not because evil people were involved, or because anyone wanted these particular outcomes; that is simply the nature of very powerful institutions. What kind of science gets done is now determined by a political process. The political nature of the process is partially masked by the fact that scientists themselves participate—those who subscribe to the prevailing views. But even these leaders in the field—as reviewers or center directors—have been known to chafe under the restrictions of the orthodoxy or to bemoan a shift in funding priorities with a new administration. The monolithic structure also pretty well shuts out unconventional, innovative research. Scientific progress, like biological evolution, depends on diversity; but that richness is lost when majority views are magnified a thousandfold. That's the crux of our problem today: The overwhelming dominance of federal funding makes it extremely difficult to do any but government-approved research using government-approved methods.

It was very interesting to me that one compelling reason for separation of science and state finally became apparent to many people just a few years ago, with the refusal of the American Psychological Association to take a clear stand against participating in torture. Their first statement said something like, "Psychologists participating in torture have an obligation to ensure that they are adhering to the highest ethical standards." There was enough of a squawk about that that the APA established a panel to study this very difficult issue and make recommendations—and allowed the Department of Defense to choose 6 of the 9 members. I think it's obvious enough to everyone what's going on: When you've made yourself so totally dependent on the government for funding, you can't afford to be critical of its policies. It's impressive how quickly we've come to such a stark illustration of that principle, and to such a flagrant display of abject dependence.

Of the various conservative institutions within the profession, we may consider briefly the role of publication, which exhibits on two levels the depersonalization of authority discussed below (p. 418): Editors rely on the judgment of their reviewers, and department chairs and tenure committees, relying on the editors' decisions, are spared having to examine people's work to assess their competence. The whole system is so tight that when, in the name of science, editors of certain journals accept only articles reporting statistically significant results (Melton, 1962), and when, also in the name of science, university faculties accredit only research published in these journals, then statistical significance is made an objective, impersonal criterion for decisions about *employment*. Science accordingly loses to politics as finding an affirmative answer to a research question becomes more important than finding the correct answer.

Reflection on the conservatism of the peer review process was prompted by the research of Peters and Ceci (1982), who selected 12 widely cited articles, changed the authors' names and affiliations to less prestigious ones (e.g., Tri-Valley Center for Human Potential), and resubmitted them to the same journals which had published them within the last 18–32 months. Only 3 of the 12 were recognized as plagiarism; of the remaining 9, 8 were rejected. The ethics of their research provided an effective distractor from their findings, but some of the contributors to the ensuing discussion, like Horrobin (1982), concluded, among other things, that the peer review process was severely biased against innovation, even without any conscious intent: Only established authorities can qualify as reviewers, and it is precisely their theories that stand to be challenged by new ideas. Redner (1987), who cynically concludes that "One of the roles of journals almost appears to be to sift out and reject really original contributions" (p. 178), discerns a kind of Gresham's law of publication: "Where too much of the work being published is standard and pedestrian, it becomes even more difficult for any unusual or really original work to gain recognition and acceptance" (p. 174). He goes so far as to exhort us to practice what Michael Thompson (1979), in his *Rubbish Theory*, calls monster conservation: to attend particularly to those ideas we reject as rubbish or as monstrous. Like rare species, we can never tell when they may become valuable.

Perhaps the oddest thing about the present situation, however, is that opposition to change is not so monolithic or entrenched as this account would make it sound. Possibly for as long as it has existed, certainly for the last few decades, the profession of academic psychology has manifested a curious ambivalence toward itself. Privately, very many psychologists are quite willing to acknowledge that much of psychological research has very little value, either theoretically or practically; many will also express doubts about quantitative methodology. Such opinions are still "politically incorrect," however, and to voice them publicly is to invite almost universal attack, or neglect. A certain groundwork has nevertheless been prepared for change, and once it becomes politically acceptable to acknowledge fundamental flaws in the traditional research enterprise, change could come quickly. In the meantime, the debate will tend to be obscured by these essentially political issues masquerading as quite different concerns.

A more benign, though still not very exciting, obstacle to change could be tagged as a lack of imagination—or, more specifically, a failure to appreciate that, as the methodology changes, so do the questions. Part of our problem now is simply that we imagine methodology moving on while we are left with the same old questions. If we drop our methodological shackles, we can presumably find ways of answering questions that are actually of interest to us. It has only been three or four generations of psychologists who have been trained to think in terms of hypothetical infinite populations, and I doubt that that way of framing questions struck any of us as natural when we first encountered it. Already by this point, however, many psychologists find it difficult to imagine asking questions in any other terms.

Kempthorne (1976) inadvertently illustrates how easily we are trapped by habit into thinking that statistical inference is the only way empirical questions can be answered. He offers the following challenge, strongly reminiscent of the famous Lanarkshire milk experiment (Student, 1931):

Someone asks: “Does a daily supplement of 8 ounces of orange juice improve the growth of children of London, Ontario?” If you deny that this is a reasonably posed problem I must part company with you. I will assume you admit it as meriting consideration. . . . I leave the problem with the reader. The type of problem is critical to the hopes and aims of humanity and I have yet to see a philosopher address it. (p. 305)

Interestingly, Kempthorne himself indicated a possible answer to his question in an earlier paper (1972), where he observed that, no matter what statistical manipulations we perform on our data, in the last analysis the prescription can only be to look at them. The judgment is apt to be difficult and “subjective,” unless orange juice is very cheap and the difference very large. If we had an instrument, such as a significance test, which would give us a sharp, unambiguous answer in such a case, we should properly distrust it, and live with the ambiguity. “Living with the ambiguity” may mean implementing the orange juice program, or not; but, in either case, it means accepting that the experimental data themselves may not strongly support one decision over the other.

Kempthorne seems to imagine that his question involves extreme sophistication for its answer, and, given the voluminous debate over statistical inference, one can perhaps see why. Not that the question lacks sophistication: Quite possibly a nutritional biologist would be of help to us in answering the question. But who would ever have thought we needed a *philosopher*?

In some respects, in short, we may have made methodology a greater obstacle than it needs to be for psychology. If we can shed the habit of thinking in terms of statistical inference and concentrate instead on what it is we want to know, we might allow methodology to recede to a more properly supporting role.

11.2.2 Possible Paths for Quantitative Research

Possible scenarios for the near future of psychological research, based on current trends, can be ordered according to the radicalness of their departure from the status quo. In the “closest possible world,” as philosophers would say, we might see psychology simply shifting to statistical tests resting on a somewhat more secure base. Where they are applicable, for example, randomization tests avoid many of the problems discussed in Chap. 9: The random process, assignment rather than sampling, is real rather than hypothetical, and so the reference class for the significance level, a permutation distribution rather than a sampling distribution, has correspondingly a more determinate meaning. The treatment of power, however, remains rather obscure, and may have to be left informal. Likelihood tests avoid the problems associated both with error probabilities and with Bayesian priors, but the stipulation of point alternatives will prove as unattractive a feature here as in the traditional theory. Either of these approaches can be employed, in a decision-making context, in a way that need make no reference to statistical inference. By the same token, the Neyman-Pearson theory retains its validity in any such context involving repetitive sampling with a need to control long-run error probabilities. But all of these will have at best a rather limited applicability to basic psychological research. Bootstrapping—just the thing for those who absolutely cannot live without a p value—would appear to have a particularly bright future, despite its exiguous rationale, if the process of its becoming known can outrun the growing disaffection with statistical inference in general.

A broader potential change is one which would retain the focus on quantitative research, but without the apparatus of statistical inference or decision. This approach has an established precedent in the research of Skinnerian behavior analysts. Although randomization tests have been used in single-case designs, Sidman (1960) argued eloquently for evaluation of trends by professional judgment and against any mechanical procedure which failed to take into account the full context of knowledge available.⁵ The simplest way of dispensing with significance tests is subtle: just not to perform them. I note with satisfaction that this approach has been taken by Gerd Gigerenzer (Gigerenzer & Hug, 1992), a thoughtful critic of significance testing himself. His theoretical nemesis Amos Tversky, scarcely to be outdone, has published an article using only a sign test to bolster one of the findings (Griffin & Tversky, 1992).

Interestingly, certain developments already occurring in the field now make this the most likely shift in the prehensile future. I refer to the increasing attention being given to the related concepts of power and effect size, due, more than anything else,

⁵ Park, Marascuilo, and Gaylord-Ross (1990) compared visual inspection by professional behavior analysts with randomization tests for an AB design, with somewhat inconclusive results. As with clinical versus statistical prediction, the comparison was somewhat unfair, in that judges did not have the full context of information that would normally be available to them. Agreement between judges and the significance tests was better than chance (evidently a real question in this case), though the judges were more conservative.

to the work of Jacob Cohen (1962, 1977) and of William Hays (1963). There are at least four reasons why power had previously been neglected. (a) One is that dealing with noncentrality parameters adds an increment to the difficulty of the mathematics. (b) More importantly, because power tables and charts were unwieldy, they were included in only the most comprehensive textbooks, like Winer (1971); it was idle for other books to enter into detailed discussions when the lack of tables precluded a numerical answer. (c) A far more important barrier, I suspect, was the fact that power calculations require the specification of a point alternative (or a set of them). With no established conventions as guidelines, stipulation of an effect size, at least in any research that wasn't simply a replication, was a matter of guesswork at best. Allowing the personal or subjective elements to come so directly to the fore defeated the main purpose of significance testing, which was to replace such judgments with automatic criteria. (d) But perhaps the most important, as well as the simplest, explanation is that pointed out by Oakes (1986): Given the Bayesian interpretation which most psychologists place on significance levels, power becomes a superfluous concept. If we think we already know the probability that the alternative hypothesis is true, why would we care about the hypothetical probability of our results on the assumption that it was true?

Cohen's contributions were threefold. First, he called attention with his 1962 paper to the fact that the estimated power of most psychological experiments (those reported in the *Journal of Abnormal and Social Psychology*) was very low.⁶ By itself, that finding would have had little impact. Second, by isolating sample size from noncentrality parameters, he was able to organize tables in a more useful fashion, and his *Statistical Power Analysis for the Behavioral Sciences* has become an indispensable reference. Most importantly, he introduced some conventions for small, medium, and large effects. Cohen was appropriately modest in putting forth his proposal for the conventions, but it hardly mattered: He had taken the first step, and it was unlikely anyone would challenge his particular definitions. With the quick acceptance of these conventions, investigators could now say that they were testing for medium-size effects (that appears to be a popular choice) without having to defend any particular numbers. Shaver (1993) indeed laments that Cohen's definitions are already becoming as conventionalized as the 0.05 and 0.01 significance levels. (He also thinks they should be called result sizes, to avoid the causal implication of *effect*—though *result* is scarcely more innocent of causal implication.)

One predictable, but unfortunate, consequence of institutionalizing somebody's particular arbitrary decision is that we quickly forget that there are any alternatives. In this case, it is often overlooked that we are under no obligation to define effects in statistical terms. Jacobson and Truax (1991), for example, criticize reliance on effect size on the same ground as significance testing, namely that a large (or significant) effect may not be clinically important. But their argument rests on a statistical definition, like Cohen's, of what constitutes a large effect. I don't think Cohen

⁶Sedlmeier and Gigerenzer (1989) reported that the power of psychological studies remained unchanged in the 25 years after Cohen's article was published, the increase in sample sizes having been canceled exactly by the practice of adjusting α levels for multiple comparisons.

himself would approve the substitution of his conventions for values derived from theory or practice.

As a result of Cohen's work, power analysis is becoming much more widespread and routine than it was even a few years ago. And, with or without conventions for small, medium, and large effects, it forces us to attend to effect sizes. As it happened, just about the same time that Cohen's article appeared, Hays published the first edition of his textbook. Having been influenced by the Bayesians, he had a better understanding than most psychologists of the issues of significance testing, and one of his contributions was the formulation of a new index, ω^2 , of effect size. The attention given to this concept in a widely used text prompted a growing, and controversial, literature on effect size. There have been at least three points of confusion and controversy. (a) Murray and Dosser (1987) point out that population and sample measures of effect size are often confused. Part of the reason is no doubt the early use of a Greek letter, η^2 , for a sample statistic, before the convention about Greek letters was well established. To try to avoid this confusion, I add a caret here. (b) Despite the early clarification by Jacob Cohen (1973), confusion persists between $\hat{\eta}^2$ and partial $\hat{\eta}^2$. In the context of a one-way ANOVA, where it was first defined, $\hat{\eta}^2$ refers unambiguously to the ratio of the sum of squares between groups to the total sum of squares—a simple proportion of variance accounted for. In multifactor designs, however, it is common—though by no means mandatory—to report the ratio $SS_{\text{effect}}/(SS_{\text{effect}} + SS_{\text{error}})$, which partials out other factors in the design. But the distinction is not well understood; even the excellent textbook by Judd and McClelland (1989), based on $\hat{\eta}^2$, slips up in one of the exercises (11.1, p. A-77), giving $\hat{\eta}^2$ instead of the partial $\hat{\eta}^2$ that would be yielded by their formula. (c) Effect sizes embody the strength of evidence against the null hypothesis, but they omit the weight of evidence, as manifested in sample size. A major criticism of significance testing is just that it bundles strength and weight of evidence indistinguishably together; but Levin (1993) is quite correct to point out that naked effect sizes can be no less misleading: His example of 2 successes in 2 Bernoulli trials (effect size = 1) recalls E. Bright Wilson's observation (Jeffreys, 1939/1961) that if a single toss of a coin turned up heads, the maximum likelihood inference would be that the coin is two-headed. This neglect of weight of evidence has led some commentators (e.g., Murray & Dosser, 1987) to argue that $\hat{\eta}^2$ is uninterpretable without its sampling distribution (which is beta)—in other words, without a p value—despite its straightforward definition as a proportion of the total (or partial) sum of squares. Murray and Dosser also argue that the dependence of $\hat{\eta}^2$ on sample size prevents comparison across studies, despite the fact that p values are much more sensitive to sample size, and less meaningfully comparable. The recent popularity of meta-analysis has focused further attention on the concept of effect size, as Chow (1988) observes. The use of significance tests both on effect sizes and in meta-analysis—as if it were meaningful to speak of a “population” of studies which the meta-analysis has randomly sampled—provides one clear indication, incidentally, how addicted psychologists are to that practice.

I think the impact of these developments with respect to effect size, however, is likely to be more far-reaching than might be suspected at this point. Once we have

gone so far as to specify the size of effect that we would regard as worth detecting or that we believe exists, it is hard not to notice that we have specified precisely what we were presumably interested in all along. If our theory suggests, for example, a 10-point difference between means, or half a standard deviation difference, or if on other grounds that is what we would take to be a meaningful or important difference, then we can just look and see whether that's what we have. Our data set, naturally, must be large enough for credibility, both in the sense of sheer weight of evidence and in the sense of representativeness of whatever aggregate to which we want to make a nonstatistical generalization. (Representativeness, it should be noted, is typically easier to justify than randomness.) At present, the focus is naturally still on significance levels, with effect sizes regarded as peripheral and perhaps even a nuisance. Over time, however, those positions could shift, so that effect sizes came to be regarded as primary, with p values becoming incidental or perhaps dropping out altogether. There is no guarantee that history will go that way; the point is only that such a shift wouldn't involve any actual change in methodology, so that it might indeed evolve simply through use.

A certain methodological shift might well then *follow* the shift in practice, as Levi (1949) observed with respect to legal reasoning (Chap. 10.) The hypothetical apparatus of random sampling theory could be retained, so that we treated our result as a sample estimate of a population value. And it would have to be retained so long as significance tests were going to be performed. But in time investigators might well come to think of their hypotheses as pertaining to the observations they were actually about to make, rather than to a hypothetical population that was inherently unobservable. The shift in practice here would again be subtle, but the implications for methodology would be momentous: that would be the end of statistical inference, or any attempt at it. A rather plausible scenario, in my view, could thus result in the abandonment of significance testing and related procedures with yet only shifts in practice subtle enough that they might almost pass unnoticed.

11.2.3 *The Ambivalent Promise of Qualitative Methods*

As more and more psychologists are becoming disenchanted with the strategy of “seeking safety in numbers,” to use Wartofsky’s (1968, p. 303) double-entendre, the last six decades have seen increasing interest in qualitative methodologies. For various reasons, however, I think those who are turning hopefully to the burgeoning qualitative literature are likely to be frustrated. In the first place, what qualitative methods have to offer is a trade-off. Psychologists have traditionally elected to try to reduce the research *question* to the point where the answer dropped out in the reassuringly crystalline form of numbers, but qualitative observations, such as interview protocols, leave us with the well-known problem of reducing the *data*. This potential for a multiplicity of interpretations is what is meant by the richness of qualitative data. But in fact we tend to look upon rich data with much the same ambivalence as we do rich people and rich desserts: We can keep our research

question closer to our actual interests but are obliged to defend our particular interpretation of the answer. Quantitative methods are commonly perceived as reducing science ideally to meter reading; the standardization of observers makes them appear interchangeable, and the act of observation impersonal, and so yields incidentally a satisfying fit with American democratic ideology. Qualitative methods, on the other hand, notoriously spotlight the qualifications of the investigator: The quality of the research depends conspicuously on our creativity, sensitivity, and preparation in general, with the disquieting consequence that not all qualitative investigators are equally good.

Secondly, a large part of the literature on qualitative methods comprises merely critiques of traditional positivist methodology and philosophy of science. Many of the contributions to the recent *Handbook of Qualitative Research* by Denzin and Lincoln (1994), for example, fall into this category. Important as such critiques are, their points are by now largely made, and they don't carry us very far toward alternative ways of doing science. Some contributions, like grounded theory (e.g., Strauss & Corbin, 1990), do look more like what we would expect in a methodology; yet it is hard to escape the impression that the further these procedures press toward codification, the more dubious they become. It may be telling that in the best qualitative research the methodology supposedly used is often invisible in the finished work. Readers of Leigh Star's *Regions of the Mind* (1989), for example, would not easily infer that the dissertation on which it was based used grounded theory.

Part of what we are up against here is just the old question of the intrinsic codifiability of research procedures. The positivists, like Reichenbach, drew a sharp distinction, as is well-known, between two phases of scientific work, discovery and verification, because they found only the latter susceptible to codification.⁷ Glaser and Strauss (1967) developed grounded theory specifically to fill the silence of positivist methodology on scientific discovery, yet they retained the premise of science as a rule-following activity. One predictable result is then that, in the limit, where

⁷Verification, in contrast to what the name suggests, is not a process of making true, but of making sure, and hence is not so much something we do to our hypotheses, as something we do to ourselves. It amounts, in one way or another, to looking to see that what we have seen before can be seen again; and a method, once identified, can be used repetitively, until we are satisfied. Discovery, on the other hand, is seeing something for the first time. Rules for discovery are thus hard to come by, because they can only be used once; then they are no longer rules for discovery. Methods for discovery elude specification, moreover, because the requisite attitude is one of receptivity to inspiration; if we followed a rule, we would not call what happened an inspiration, or the result a discovery.

In these terms, the distinction is familiar enough, but it can very easily be overdrawn (Feyerabend, 1975). The processes of discovery and verification may best be thought of, perhaps, as marking more or less incidentally two complementary aspects of ongoing inquiry. The phase of verification represents the formalizable aspect of scientific inquiry and rule-following activity; the process of discovery is unformalizable and represents, as it were, the rule-breaking aspect. Both processes go on in alternation, almost concurrently. To find the distinction actually becoming sharp in practice, Feyerabend suggests, is indicative of a temporary stasis in research; we might go on to say that our tendency to find the sharp distinction plausible is indicative, perhaps, of the static quality of our conception of research.

these methodologies lend themselves to mechanization in computer algorithms, they tend to reproduce the context-stripping features of typical quantitative methodology. As a further, special consequence, findings in such research will tend to reflect, generally in nonobvious ways, the basic assumptions built into the structure of the software (Richards & Richards, 1994).

As a glance at the contributors to the Denzin and Lincoln *Handbook* (1994) will suggest, most of the developments in qualitative methodology have come from outside psychology, notably from sociology and anthropology, which had a harder time all along capitalizing on quantification. The distinctive contribution of psychology in this area lies in what it has done to phenomenology, just as its distinctive contribution to quantitative methodology has been factor analysis. It is tempting to characterize these two methods as about equally well-founded. In the 1960s, when the search for an alternative philosophy of science acquired a critical momentum, phenomenology was the most obvious contender (Wann, 1964). To make research under such a paradigm more acceptable within the positivist hegemony, however, pioneers like Giorgi (1985) made phenomenological psychology a study of what *other people* think; and American students ever since have supposed that phenomenological research meant doing interviews. Giorgi's efforts to articulate a specific methodology, moreover, evidently put him in an embarrassing position like that of authors of successful self-help books: People—in this case, psychology graduate students—are so desperate for someone to tell them what to do that anyone who has any suggestions to offer soon finds that she or he has to beat off disciples with a stick to keep from being elevated into a guru. The dutifully performed rituals of chopping narratives into “meaning units” and the like bear precious little resemblance to phenomenological method as articulated by, say, Spiegelberg (1982). But the press for practical rules of procedure is making it depressingly commonplace to read dissertation proposals whose method chapters witlessly begin: “Step 1. I will set aside all presuppositions . . .”

Quite beyond the disappointingly mechanistic flavor of some of the analytic procedures of such research, it is often hard to discern in what respect it is supposed to qualify as phenomenological. Many self-identified phenomenological studies are recognizable chiefly through their abundant use of hyphens, as though to suggest that the authors' thought were too subtle to survive translation from German. In part the problem reflects simply the deterioration in the meaning of the word: *Phenomenological* is now commonly used to refer to any nonbehavioristic research; and, behaviorism being characterized by its exclusion of experience, it is also used merely as a synonym of *experiential*. The central concept of essence seems rarely to be grasped at all. The search for structural invariants in the flux and diversity of human experience may be an odd one, but a statistical interpretation of essences in terms of typicality (or less)—by those who claim to reject statistical thinking—is embarrassing. In the first American psychology dissertation to identify itself as phenomenological, in 1958, Adrian van Kaam took as an explicit aim the identification of the necessary and sufficient constituents of the experience of “feeling really understood”; yet the list of constituents he derived from 365 (!) written narratives applied to as few as 64% of his respondents. Some of them, like “feeling satisfied,”

were also of questionable sufficiency. More disturbingly, von Eckartsberg (1986), presenting van Kaam's study in detail, makes no criticism of these core lapses of logic.

Significantly, the originators of phenomenology themselves offer little help; one searches their work in vain for a credible application of phenomenological method. Evidently, the closest that Husserl himself came was in the second volume of his *Ideas* (1989), on the phenomenology of constitution; but his a priori partitioning there of the world into material, animal, and spiritual begged the very questions he was so meticulously addressing, and he withheld publication during his lifetime. And though phenomenological analysis can in principle be conducted admirably through fiction (Ihde, 1977), there is no way the novels of Sartre, with their unmistakable personal coloration, could be construed as phenomenological explorations.

That qualitative methods, in general, are such an easy target reflects, I believe, our unrealistic expectations of them. The very high—spuriously high—degree of codification of quantitative methodology in psychology has led us to assume that this is what methods are supposed to look like: recipes that can be followed mechanically, ideally without any judgment or independent thought. Consequently, I think qualitative methodology is already beginning to function as a kind of transitional object, easing us ever so slowly away from the predefined codification of quantitative research, until we can face the world on our own.⁸

11.2.4 *Toward Deeper Obstacles to Change*

But I remember being asked (to my surprise), when I began lecturing to psychologists on the history of statistical inference: “Suppose we grant everything you say up to this point; then, if we can't do statistical inference, how are we supposed to analyze our data?” (How on earth can anyone cook without recipes?) The answer that I really wanted to give is, naturally, “Why do you insist that I—or anyone else—tell you how to analyze your data?”

For a slightly more elaborated answer, I could offer an analogy. Suppose you were to challenge the Ten Commandments as the last word on ethics (it has already

⁸ It is commonly observed that, of the psychologists who would be on almost anyone's short list of “greats,” most are Europeans who were working before World War II (and the subsequent importation of American research methodology), and none of them relied on existing formal methodology. Skinner may be the American most likely to make such a list, but he had as little use for statistical inference as Piaget.

Such work is sometimes belittled on the ground that, without statistical inference, it remains speculative (and therefore, presumably, valueless). Freudian theory, in particular, has been charged, with some justification, with invulnerability. But if a theory is susceptible to correction, then it is still saying something. I am not thinking of the silly studies by American psychologists claiming to “refute” Piaget by showing that American children lag far behind the Swiss in their cognitive development; I am thinking of such work as Gilligan's (1982) critique of Kohlberg's (1969) theory, and the Dreyfuses' more profound critique of both of them (Dreyfus & Dreyfus, 1992).

been done). In the unlikely event you succeeded in planting any doubt in a fundamentalist, you would be sure to be asked for a new set of rules: “If these aren’t the correct rules to live by, which ones are?” It would be exceedingly difficult to get across the idea that good conduct is not just a matter of following rules. In fact, rule following may have nothing at all to do with virtue. By the same token, I would submit, following Feyerabend (1975), that neither is good science a matter of following rules. Feyerabend indeed insists, as was noted in the previous chapter, that the real scientific advances have come from *breaking* the rules, and that the only rule not guaranteed to inhibit progress is “Anything goes.” If it is tempting to dismiss Feyerabend (1975), the first philosopher of note explicitly to advocate epistemological anarchism, as a firebrand, consider the more sedate musing of the statistician Lucien Le Cam (1977):

If inference is what we think it is, the only precept or theory which seems relevant is the following: “Do the best you can.” This may be taxing for the old noodle, but even the authority of Aristotle is not an acceptable substitute. (p. 134)

But I have discovered by now that my audiences are not very satisfied with that answer, either.

Epistemological anarchy raises the same specter of chaos and irrationality as political anarchy, and for essentially the same reasons. So, it will be useful to consider why we have ever supposed we needed to be governed by rules in either domain. If we have not perceived ourselves as rabid fundamentalists with respect to methodology (or social organization), it is just because such fundamentalism has been the mainstream until recently.

If statistical inference has represented a pinnacle of positivist methodology, the assault on it is yet but one piece of the undermining of foundationalism spreading throughout the culture at the end of the last century. That century was distinguished from the beginning by radical questioning (e.g., dadaism), but the horrors of two world wars and the Holocaust had the countervailing effect of intensifying the search for foundations—a project which is exhausting itself only in the last few decades. Much of the impetus for change has been coming, significantly, from groups which have been oppressed by this culture, including women (e.g., Diamond & Quinby, 1988; Harding, 1986, 1991; Jaggar & Bordo, 1989; Nicholson, 1990). Certain fault lines are thus built in (Berman, 1989, makes much of this pun); in particular, we can expect polarization between those, on the one hand, seeking to subvert the Church of Reason which has been oppressing them and, on the other, the antidisestablishmentarians looking for traditional supports—and pointing accusingly around them to increased debauchery. The recent debate between Kenneth Gergen (1994) and Brewster Smith (1994) is a mild example.

It is for these reasons, William Irwin Thompson (personal communication, February 4, 1993) suggests, that we are especially sensitive at this point to the issues of the seventeenth century—except that the foundations frantically erected in that crisis are the ones collapsing around us now. The seventeenth century was “saved” by “the flight to objectivity,” to use Bordo’s (1987) phrase; the story of the twentieth century might thus be called, following a suggestion by Gail Hornstein, *The Crash*

of the *Flight to Objectivity*. In response to the collapse of religious authority, we made a religion of science; it is not clear where we can turn next in our quest for authorization.

11.3 The Philosophical, Social, and Psychological Context for the Emergence of a New Epistemology

The fear of anarchy and relativism implies, paradoxically, feelings of both impotence and omnipotence. Feelings of impotence are contained in the implication that we would be helpless without rules or other authority to tell us what to do. As such, these feelings provide psychological support for the various practices of observer exclusion, which helps to disguise our personal responsibility (the Eichmann approach to research). Feelings of omnipotence are contained in the implication that nothing outside the self matters, that we need external checks to constrain us. They are thus relevant in understanding the perversions of control in modern science.

From the presence of feelings of both impotence and omnipotence, the least we can conclude is deficiencies in self-regulation. In fact, I shall argue that the post-modern dilemma—the perception that we must choose between the absolutes of foundationalism and the anarchy of relativism, and the inability to embrace either—results from an all-or-none conception of control not unlike the alcoholic's. If we want to understand the extraordinary pervasiveness of problems in control, we may need to look no farther than child-rearing practices. Trying to control children—trying to get them to do what we want—through an external system of rewards and punishments is all but universal. With children, as with pets, we can achieve a superficial success as long as we can maintain control of the food supply, and otherwise physically overpower them. But Powers (1973) has shown clearly why such a system cannot work in principle, and can only lead to violence in the long run. In general, practices which exacerbate the powerlessness in childhood can lead either to feelings of impotence and despair or—if powerlessness is identified as a limitation of *childhood*—to omnipotence, to the assumption that adults can do anything.

The first task of this section will be restoration of epistemological license to the individual knower, but that will still leave us taking drastically insufficient account of the social aspect of knowing. Indeed, it is possible to interpret our attachment to rules and external authority as a disappointment-hardened substitute for a lost capacity for conviviality.

The theories I shall adduce, by Dinnerstein (1976) and Benjamin (1988), will help in understanding not only the substance of the issues, but, perhaps more importantly, why we confront so many false dichotomies in our culture: Atomistic individualism versus mindless collectivism, dominance versus submission, totalitarianism versus nihilism. Although the primary concern here is with epistemology, it will be illuminating to notice the same issues in politics. All of these,

needless to say, are topics for books in their own right; but, since the questions rise before us, I want at least to indicate what kind of answers may be possible.

I might as well acknowledge also at the outset some problems with the analysis that follows. One has to do with the intrinsic inadequacy of ontogenetic explanations for cultural phenomena: universal explanans, particular explanandum. The principles I shall adduce from Dinnerstein and Benjamin are generic; yet not all Western cultures have, for example, made such fetishistic use of significance testing, or even of quantification, in psychological research as has the U.S. The very different course of psychological research in Germany—a country which reportedly (Stone, 1977) went straight from wet nursing to bottle feeding—presumably had little to do with child-rearing practices and much to do with institutional sources of support (Danziger, 1991). The combination of social unrest from massive immigration into the United States around the turn of the twentieth century and a pragmatic “cult of efficiency” (Callahan, 1962) modeled on the wonders of mass manufacture had created an opportunity for psychologists to market themselves as the technologists of the progressive-utopian reform movement; and the pseudoscience of significance testing was particularly useful for dazzling the nonspecialists in charge of funding. In Germany, on the other hand, psychology remained closely tied to philosophy until World War II. Developments since that time, however, may be relevant to the arguments in this chapter. Spiegelberg (1982), speaking of phenomenology, notes how ironic it was that France, after World War I, should be smitten with a philosophy which bore all the worst earmarks of German thought; it is hardly less ironic that Germany, after World War II, should so zealously have begun adopting American psychological research practices. Perhaps psychodynamic explanations are not *completely* irrelevant.

A second problem in using ontogenetic theories to understand philosophical issues is the implication that historical changes in philosophy should be paralleled by changes in child-rearing practices. That is a strong claim, more decisive evidence for which will have to await future scholarship. It is true that the latter half of the seventeenth century and the first half of the eighteenth constituted, at least in England, a period of relative humanity and permissiveness, following a century or so which Stone (1977, p. 217) characterizes as “the great age of the whip.” The collapse of Puritanism (and the flight of its followers to Massachusetts) led to a brief period of antinomianism, evidently experienced as a kind of cultural adolescence, with philosophical order being restored in part by the Lockean doctrine of natural law. The Lockean theory of tabula rasa also helped to undermine the concept of original sin, allowing a more humane treatment of children. Stone links the development of a new personality, characterized by a “steep gradient of affect,” to child-rearing practices based on trust rather than distrust.

Stone (1977) also emphasizes, on the other hand, that the history of childhood in the last four centuries has been anything but a steady progression toward enlightened child-rearing practices, but, rather, alternating waves of brutality and humanity, with sharp differences across social classes. But the philosophical and political changes that were instigated in this century-long window of relative tolerance lasted well beyond it, and some of the psychological advances may have, too. And

philosophy is largely the product of the upper classes, of those with time on their hands, and these have tended to be the more liberal in their attitudes toward children. Wealth brings certain constraints of its own—marriage partners are more carefully scrutinized, for example—but John Wesley’s sermons on breaking the will of the child found a more receptive audience among the poor. The evidence I can find, in short, at least does not controvert a link between the child-rearing practices and the philosophy of a culture. Now to some details of possible such links.

11.3.1 *Empty Self*⁹

In the first place, a clinging to rules suggests not only a deficiency of self-governance; a lack of reliance on our own perceptions and judgments for guidance suggests an alienation from our own experience, or a distrust or rejection of it. Such an estrangement from our own experience is fundamental to what David Shapiro (1965) calls the obsessive-compulsive cognitive style. The consequence of estrangement from our experience—which entails, in crucial respects, a loss of connection to the world—is a need for strong structures of control and authority; and the obsessive-compulsive style is indeed characterized by devotion to ritual and to rule-following. So in this respect, the pathogenesis of contemporary psychological research methodology can be construed as a cultural obsessive-compulsive disorder.

In characterizing the obsessive-compulsive style, Shapiro (1965) emphasizes an impoverishment of inner experience and a correspondingly reduced sense of self as volitional agent. An unforgettable example of his is the lawyer who selected his wardrobe each morning with the aid of a color wheel. It was obviously an alien idea for the lawyer to consult his own experience, just to look and see what looked good; lacking any impressions of his own, he had to rely on an external set of rules embodied in the color wheel. With the sense of self as incapable of making choices, guidance and control are felt as external, and the reliance on rules leads to the excessive rigidity characteristic of obsessives.

⁹I am aware that the Buddhist concept of empty self is regarded as a kind of ideal of mental and philosophical health (e.g., Rosenbaum & Dyckman, 1995; Varela, Thompson, & Rosch, 1991), and I don’t necessarily disagree with their view. I see it as a privileged position, however, of those who have already constructed a secure self and world, and then have the luxury of playing with alternative constructions. Otherwise, the rhetoric of empty selves can make it sound as though multiple personality disorders were the epitome of enlightenment.

There is also ambiguity in the scope of the concept of self in this literature. The experience of taking different “selves”—i.e., roles—in different social contexts is familiar enough; but if people respond to us as though we were literally different selves, then we are apt to feel simply misunderstood, unseen.

In comparison with the exalted concepts of Buddhism, my own headings *Empty Self* and *Empty World* are somewhat overblown, referring to something more like a garden-variety cultural neurosis.

The hysterical style, on the other hand, is essentially impressionistic. It proceeds from an attitude of receptivity, and conduces to a more physiognomic perception; and it finds a more natural use of language in expression than in representation.¹⁰ Just as the potential cognitive advantage of the obsessive-compulsive style, subject-object distance, has its defensive possibilities, so the hysterical style has its defensive uses in blurring the meaning and blunting the impact of events.

As pure types, neither of these, even in Shapiro's attenuated form, is epistemologically adequate in itself. It is obvious enough that scientific knowledge, in any conception, depends on the capacity for prescission from immediately given concretes and for the achievement of distance both from things and from ourselves in the process of knowing. It is equally true that when we block our sensitivity, we not only impair our primary source of information about the world and thereby our experience of efficacy in dealing with it; we also leave conceptual thought with an impoverished base, and an extra burden in the task of understanding. In cutting off the organic roots of our cognition, we reduce ourselves to the dry husks that Barbara Branden (1962) used to call psycho-epistemological Platonists. The problem for modern science is again its one-sidedness, the rejection of the hysterical mode of knowing and corresponding hypertrophy of the obsessive-compulsive mode.

It will have been obvious in the foregoing characterization of cognitive styles that that distinction is aligned with and reinforced by gender. The qualities that are rejected by the obsessive-compulsive—sensitivity, vulnerability, receptivity—also happen to be culturally identified as feminine, so that their rejection, disabling as it is, tends to feel right, to many women as well as to most men. What is relevant is not the statistical fact that the hysterical style is more common in women and the obsessive-compulsive among men, but that the hysterical mode of knowing is devalued, suppressed, and denied, by men and women alike, while the masculine-identified obsessive-compulsive mode is exalted and hypertrophied in its development. The relationship between gender identifications and science has been subject to a gratifying explosion of scholarship starting in the 1980s. To my mind one of the most profound and provocative explanations of such gender patterns lies in the seminal work of Dorothy Dinnerstein (1976). She articulated the implications of her theory for political authoritarianism, but did not extend it to epistemology; so I want briefly to recapitulate her theory, which pivots on the fact that child-rearing has always been a monopoly of women, and point to the implications for cognition.

Many observers, of different orientations, have interpreted the human history of repeated subjugation to authority, plausibly, as an “escape from freedom.”

¹⁰As was noted in Chap. 2, the greater self-world differentiation implicit in the new status of language in the seventeenth century led to a more objective use of language, for representation. As Foucault (1973) observes, the functions of language at that time split, so that more expressive uses have since been relegated to literature, to poetry, and dispatched from science, from knowledge. The split is less thoroughgoing than the others discussed here, with scientific discourse reaching only a modest pinnacle in the approved journalese of what in psychology is ironically called “the literature.” One reason is presumably that it is still too hard to talk like Bertrand Russell in the *Principia Mathematica*; we are too obviously dependent on ordinary language to succeed in suppressing it—at least until a Fisher of words is forthcoming.

Dinnerstein, on the other hand, offers the view that it represents escape from a still greater authority. The significance of women's monopoly in child-rearing lies first in the fact that all the rage and frustration of infancy, of unmet needs and the struggle for autonomy, are directed against women.

It is obvious that we all have character traits which make us less than perfectly parental. What is not faced head-on is the fact that under present conditions woman does not share man's right to have such traits without loss of human stature, and man does not share woman's obligation to work at mastering them, at shielding others from their consequences. Woman never will have this right, nor man this obligation, until male imperfection begins to impinge on all of us when we are tiny and helpless, so that it becomes as culpable as female imperfection, as close to the original center of human grief. Only then will the harm women do be recognized as the familiar harm we all do to ourselves, not strange harm inflicted by some outside agent. And only then will men really start to take seriously the problem of curbing, taming, their own destructiveness. (Dinnerstein, 1976, pp. 237–238; original in italics)

But the split also makes it possible for us to project the different sides of a number of fundamental ambivalences onto each gender. This “solution” generally ensures that one side or the other is disowned, that each is alienated from the other and hypertrophied in its expression, and that the ambivalence itself is never recognized or dealt with, as it would have to be if both sides were represented in each of us.

One of the things we (i.e., Euro-Americans of the last four centuries) feel ambivalent about is the body. As our concrete, physical existence, it is of course the source of pain as well as pleasure. Of more relevance in the present context are its implications for cognition: In order for us to have any awareness at all, it is necessary that our consciousness have some specific form, including a particular sensory apparatus; but then that form by the same token limits our awareness. Physical existence also limits us to being in only one “place,” with its particular perspective, at a given time—though it is the dimension of time which makes it possible for us to move around, and in particular to look back on the vantage point where we formerly stood, thus achieving some form of objectivity. Still more relevantly, our bodily experience of the world includes a glandular as well as a neural component; the two systems differ in inertia, and their lack of synchrony frequently confuses our apprehension of the world.

Emotion, owing to its greater mass momentum, is unable to follow the sudden switch of ideas to a different type of logic or a new rule of the game; less nimble than thought, it tends to persist in a straight line. Ariel leads Caliban on by the nose: she jumps on a branch, he crashes into the tree. (Koestler, 1964, p. 58)

The deep identification of the body, like nature, as feminine makes it clear, on Dinnerstein's theory, why it is the bodily component of knowing that should have been subjugated to the masculine ideal of pure, ethereal mentation. It is at this point, Dinnerstein herself could well have written, that the project of healing the split between mind and body, and the project of sexual liberty, meet—with the future of science hanging in the balance.

Another thing we feel ambivalent about is the process of growing up.

Few of us ever outgrow the yearning to be guided as we were when we were children, to be told what to do, for our own good, by someone powerful who knows better and will protect us. Few of us even wholeheartedly try to outgrow it. What we do try hard to outgrow, however, is our subjugation to female power: the power on which we were dependent before we could judge, or even wonder, whether or not the one who wielded it knew better and was bossing us for our own good; the power whose protectiveness—although we once clung to it with all our might, and although it was steadier and more encompassing than any we are apt to meet again—seemed at that time both oppressive and imperfectly reliable.

Having escaped that power, or at least learned how to keep it within bounds, all but a few of us have exhausted our impulse toward autonomy: the relatively limited despotism of the father is a relief to us. (pp. 188–189)

If men, however, bore equally the burden of those infantile feelings now attached to women, then subjugation to authority—male authority—would not hold the appeal that it does for us.

If a different, apparently blameless, category of person were not temptingly available as a focus for our most stubborn childhood wish—the wish to be free and at the same time to be taken care of—we would be forced at the beginning, before our spirit was broken, to outgrow that wish and face the ultimate necessity to take care of ourselves. (p. 189; original in italics)

Dinnerstein's dynamic is certainly not the whole reason why we appeal so universally to authority, but her explanation is as applicable to epistemology and ethics as it is to politics. It was remarked in Chap. 2 that the impact of the Renaissance was a secularization of epistemological authority, and the process has continued, in the political as well as the philosophical realm, with the depersonalization of authority. It is very important to us now that authority not reside in a person but in the impersonal structure of a set of rules or a bureaucracy—despite the fact that we can scarcely count on a greater responsiveness (the IRS comes to mind as an example). The orientation to authority has remained the same, however, through a succession of objects.

The qualities of unself-consciousness and impersonality have defined our paradigm, a cognitive Eden. Supposedly the ideal of an assured, effortless existence was lost when we acquired self-awareness and individual responsibility. On the present view, however, our real sin is yielding to the temptation of an automatic, guaranteed form of knowledge. That is the source of our guilt, of our intellectual self-distrust. Under a putative ideal of cognitive innocence, undifferentiated as knowers, we are unaware of ourselves in the process of knowing, and whatever knowledge we possess in that way is impersonal and unreflective. In partaking of the Tree of Knowledge, we become individuated as knowers, and become aware of ourselves. Self-awareness is the precondition of personal, reflective knowledge, and with it comes the necessity of cognitive effort and individual responsibility.

Although, according to Keys (1972), the Chinese understood at least as long ago as the fourth century B.C. that the act of observation affects what we see, the principle was not understood in the West until the twentieth century. The methodology of self-exclusion has pervaded Western science, and psychology, ironically, most of all. Martin Orne's (1962) work on demand characteristics and Robert Rosenthal's (1963) work on the experimenter effect were primarily responsible for calling the

attention of psychologists to the principle of observer participation. A further step was taken by Forward, Canter, and Kirsch (1976) in their criticism of the deception methodology of social psychological research. They pointed out the various fallacies of self-exclusion, including the fact that the experimenter cannot ever determine if the deception has succeeded except by relying, sooner or later, on the subjects' verbal reports—just the sort of thing the deception strategy was designed to avoid. Forward et al. recommended in its stead the role-enactment method, where the subjects are invited to participate in the study as coexperimenters. They are essentially handed "scripts" and asked to act out or to predict their own behavior in the given situation. The role-playing methodology has many advantages over the deception methodology, but it is also subject to several limitations. We must be concerned about our participants' role-playing skills, about their motivation for the task, about role-participant congruence, and so on. Forward et al. stopped short of the next logical step: If we are going to bring nonprofessional people into the laboratory, train them carefully to understand a situation just the way we see it, and ask them for their reactions, why not simply play the role ourselves and eliminate the problems of the middle people?

A similar point can be made about all the studies involving the creation of a new scale to measure some psychological dimension. It is obligatory in such studies to secure a reliability coefficient by training at least one other person to see things as much as possible the way we see them, then measure the extent of agreement (or, more accurately for the typical case, covariation; see Tinsley & Weiss, 1975) in our judgments. The coefficient really measures little but the effectiveness of our instruction, and what is the point? Presumably we know better than anyone else what we want to capture in the scale, and are accordingly the best judge of how well we have succeeded. Morse (1994) points out that it is not considered obligatory to take colleagues with us to the library as a check on the reliability of our reading of the literature.

If you read a modern university textbook on, shall we say, psychology, you would think the author didn't have any experience of his own. I know that the reason given for this extraordinary omission is that, in respect of one's own experience, one is likely to be biased and therefore not "objective." But if you cannot be honest about your own experience, how the hell can you expect to be honest about anyone else's. And if you think you are likely to be "mistaken" about your own experience, how much more likely are you to be mistaken about somebody else's experience of which, by definition, you have no experience. (Keys, 1972, p. 13)

Indeed the whole elaborate apparatus of experimental and statistical manipulation seems designed to keep us from noticing that in the last analysis there can be no escape from looking and seeing for ourselves. If we merely rely on a formula or cookbook instructions, still at some primitive level we endorsed the formula as appropriate or correct. As an anonymous contributor to the Diderot *Encyclopédie* put it, "*Il faut avoir les données, mais on ne doit emprunter la solution de personne*" ("Fatalité," 1756, p. 428). Mainland (1960) cites an unnamed research worker as saying that the reason we need statistics is that "we can't all be Claude Bernards," so we need statisticians to tell us what to think. I would say that if you

need somebody to tell you what to think, you shouldn't be doing science; and I would also thank you for not voting.

We have been discouraged from accrediting our own perceptions, and from accepting the responsibility for knowing, by our entire orthodox philosophy of science. It is a philosophy of (how could Ayn Rand have missed the phrase?) epistemological altruism, based on self-distrust, and reliance on external, codified authority, for validity, meaning, guidance, structure. To use Langer's (1967) apt label, we have erected the Idols of the Laboratory as chimerical projections of what we have abandoned in ourselves. Having drunk too deeply from the Cartesian well of doubt, we are frozen in criticism and self-doubt, as our only approved intellectual activities, and have blocked off possibilities for growth and development. It has been said (Thurber, I think, once again) that in preferring the *Cogito ergo sum* to the dynamic *Non sum qualis eram*, we are putting Descartes before Horace. Cognitive growth, as Branden (1984) has emphasized for personal growth, can only start from self-acceptance.

11.3.2 Empty World

Toward reclaiming the social, or the intersubjective, aspect of knowing, I rely on the work of Jessica Benjamin (1988), who shows that it has been missing, not just from history and ontogenesis, but even, remarkably, from most modern feminist theory.

The roots of the intersubjective lie in the experience of mutual recognition. All of us need to see and to be seen.¹¹ But to be seen as a person requires the perspective of another person; the recognition must come from an other that we recognize as a self in her own right. In various ways, however, our culture has denied women the role of subjects. The ideal mother has been conceived in terms of nurturance and responsiveness, attending in every way to the child's needs; any independent needs or desires of her own constituted a limitation, a problem. Many feminists, Benjamin (1988) points out, have supported this cultural status quo in their rejection of women's sexuality. But in fact an overcompliant parent, one of which we are too thoroughly in control, deprives us of the experience of mutual recognition; the other is too literally of our own creation to constitute a subject like ourselves, and so to serve as a mirror. Some testing, and finding, of limits is thus needed to establish the alterity of the other.

¹¹ See Branden (1969) for a discussion of the latter need, which he calls the visibility principle. (In David Witter's formulation: "I'll show you yours if you'll show me mine.") It is characteristic of Branden's rationalism (at least at that time), I am inclined to say, that he would have grasped the fundamentality of the need for *psychological* visibility while overlooking what one would think was the more obvious need for *physical* visibility. The need to see and be seen as physical entities, which has been correspondingly frustrated by our culture for about the same period, would appear to have been grasped implicitly by nudists—many of whom, however, remain caught up in issues of social acceptability in our present cultural context and whose articulation of the principles involved remains therefore disappointingly muddled.

The denial of female subjectivity hardly needs to have succeeded fully in practice for its deleterious effects to be felt: It still leaves the implication that there should be no limits in principle on our power. But, it also goes by now without saying, these effects are felt differently by males and females. Girls have a harder time under the circumstances separating from the mother and establishing an autonomous identity, whereas for boys the problem is distinguishing the process of becoming a separate person from becoming a male person. When the gender confounding is unrecognized, boys grow up equating dependence with femaleness and relating to others as objects. The development of intersubjectivity is thus impaired, in different ways, for everyone. To the extent that it is realized, it is confined to the private, feminine realm, now sharply distinguished from the public. Science and politics, occupying the latter, are now pervaded by impersonality and atomistic individualism. Moreover, the sharp separation itself, as well as the characterizations of the two sides, has tended to feel right to both men and women.

The latter half of the eighteenth century, when statistical inference was being developed, is noteworthy as the period of revolution against the state, in the name of individual liberty; but Foucault (1979), and Sampson (1988) following him, suggest that something more subtle was going on than the liberation of the individual from the authority of the state: more significantly, a redefinition of the individual as an autonomous, isolated entity.¹² Although American patriotic lore continues to celebrate, at least ritually, the concept of the autonomous individual, the actual process, if these authors are right, may have been more insidious: for Sampson (1985) argues persuasively that it is an aggregate of atomistic, isolated individuals, leading a monadic existence, who most need a strong central government. Societies, on the other hand, with strong personal and group bonds can get along very well without a formal central government, and have done so (see, for example, Byock, 2001, and Friedman, 1989). Similarly, once the natural bonds between things had been dissolved, a powerful artificial device—statistical inference—was required to bring them back together under cognitive control.

The process, moreover, is an unfortunately self-reinforcing one. Charles Murray (1988) has shown (see also Code, 1992) that, just as extrinsic motivation tends to undermine intrinsic, so the introduction of external control structures tends to weaken natural social bonds. The welfare state has not (fully) replaced private charity, but we are less likely, nevertheless, to help a neighbor in need when we can assume that there is some bureaucracy that is supposed to take care of the problem. Similarly, it can be argued, with defense: When Florida restored the right of people to carry concealed weapons in 1987, it saw a drop in the murder rate over the next 7 years, during which time it climbed by a similar amount elsewhere (cf. e.g., www.lewrockwell.com/2005/03/john-lott-the-right-to-pack-heat/). In science, reliance on

¹²The actual transformation of society was hardly achieved instantaneously, of course. The atomistic social philosophy of Locke was ultimately realized not so much stipulatively through the U.S. Constitution as practically through technology: The telephone and the airplane, in particular, drastically weakened traditional ties—to place, to family, and to the community—in allowing us to relocate across distances we previously would not have considered.

formal methodologies has made the skilled judgment to which Sidman (1960) appeals seem superfluous; it is not regarded as a necessary part of the training, even for qualitative research. In all realms, as we relinquish our own responsibility for ourselves and others around us, we confront an apparent need for ever stronger controls. For solutions to social problems, it gets harder and harder to think of anything other than more federal money.

Some years ago Hilde Schlesinger told me that, in a study of the development of deaf children, she and her colleagues had asked parents to describe their concept of the ideal 9-year-old child. When they followed up by asking parents to relate some of the things they had done to try to make sure their child turned out that way, a few wise parents smiled in recognition of the trap: that the way to have an ideal child was not to try to force him or her into some preconceived mold. The point is beginning to be appreciated with respect to child-rearing, just as the paradoxical futility of trying to achieve social ideals by physically enforcing them is beginning to be appreciated, as for instance in the “wars” on drugs or poverty. The similarly self-defeating nature of our one-sided pursuit of control in science, however, has been less noticed.

Science indeed defines its goal—or, rather, *philosophers* have defined the goal of science—not as understanding, but control (and prediction, which is cognitive control). So strong are our needs for control in psychology that we typically undertake “experiments,” not to learn, but only to demonstrate what we already believe. It is now largely forgotten that *experiment* derives from the same root as *peril*, and that it once connoted adventure into the unknown and a placing of ourselves at some risk.

Logically, however, we should expect knowledge to be served by observation under conditions of least interference and control, rather than the greatest. “To observe anything in the outside world, we have to interfere with it, for example by shining a light on it. And the more sensitive it is, the more the interference changes it” (Keys, 1972, p. 30). The paradox is that the more we are in control of the knowledge process, the more subjective it becomes, the more it reflects ourselves as knowers.

Piaget (1950) has provided the best illustration of this principle, and his concept of knowledge as an equilibration of assimilation and accommodation also happens to provide a serviceable model for understanding our one-sided emphasis on control. This model of knowledge is surely the most familiar to American psychologists, but they have tended to regard it as pertaining only to children’s learning about rolling clay into sausages. It is just as applicable, of course, to their own scientific practice—so long as they are actually engaged in learning; but that practice, and its supporting theory, grossly slight the accommodative aspect.

Since Comte (1830–1842/1864), we have been accustomed to thinking of the sciences as comprising a linear array, from psychology and the human sciences at one end to mathematics and logic at the other. In Piagetian terms, the sequence can be comprehended by the poles of accommodation and assimilation, respectively. Piaget (1950) himself, however, sees the sequence as folding back on itself into a circle, from mathematics into psychology. In mathematics, the most purely constructive of the sciences, we are most “in control”—positivist philosophers of

science have even seen mathematics as purely “stipulative.” But the result is then that it is but the constructive operations of human intelligence which constrain and define the discipline, and they become the object of our study. The most “objective” form of inquiry, in short, is *ipso facto* the most “subjective.”

The “circle of sciences” is an especially useful conception because it points up the double aspect of all knowing. *Universe*, as Spencer Brown (1972) reminds us, means “one turn,” and we can get different worlds, inner and outer, depending on which way we traverse the circle.¹³ We can illustrate the principle by reflecting on a common problem in experimental design, one which has helped create a scandal with analysis of covariance (see, for example, Lord, 1969), but also arises with the simple device of matching, as in the deservedly classic study by Meadow (1967/1968, 1969).

Meadow hypothesized that deaf children of deaf parents would surpass deaf children of hearing parents in their linguistic, academic, and socioemotional development, owing to the benefits both of early (manual) language acquisition and less family trauma surrounding the birth of a deaf child. To test her hypothesis she matched 58 pairs of deaf students from deaf and hearing families on a variety of background variables, including IQ, family climate, and socioeconomic status (as indexed by father’s occupation). The implied question in the comparison is, “What would deaf children be like if they had deaf instead of hearing parents?” and that is, again, just the question that is involved in selection of appropriate controls to answer the original question. It might be argued that at some early age deaf and hearing children are still sufficiently undifferentiated to allow a plausible match, but by adulthood matching on some variables is no longer possible. There are, for example, unfortunate but obvious strong differences between deaf parents and hearing parents, as groups, on socioeconomic status. To achieve even a semblance of a match required equating skilled craftsmen among the deaf fathers with professional, managerial, clerical, and sales workers among the hearing fathers. Even with this liberal criterion, Meadow regarded 10 of the pairs as unsatisfactorily matched. In general, matching absolute levels on such a variable requires selection of the upper end of one distribution and the lower end of the other. Of necessity, then, the *relative* positions of these families in their respective contexts are different, and it is entirely reasonable to expect this latter variable (relative sociocultural position) to make a material difference in the results. The pride felt by highly successful deaf families, or the social shame felt by unsuccessful hearing parents, for example, could be expected to affect their children’s self-esteem, which was one of the dependent variables of the study.

In general, as we force equality on some characteristics of inherently different intact groups, we force inequality on others, which confounds our inferences. The problems do not arise when we have randomly assigned individuals to treatment

¹³ It may be apropos to recall that *research* comes from the Latin word *circus*, meaning circle. The related verb *circare* meant to go around or to go about; French softened it to *chercher*, and it is from *chercher* and *rechercher* that we get our English *search* and *research* (Partridge, 1966). Hence *to research* literally means to keep going around in circles.

conditions, for in the random assignment we have effectively detached the imposed treatment differences, and in that case, our additive linear model is most defensible. Mathematically, the model can still be applied to situations where we have made no actual changes in the subjects of our study, but the mathematical adjustments do not take into account how the hypothetical changes were made. It is seldom realistic to assume that characteristics of intact objects or organisms can be added or subtracted at will without impact on other characters; and the question, posed without regard for the mechanism of change, is often rather hopelessly hypothetical. As Sidman (1960) put it, “Statistical control of multiple causation is a device for manipulating the verbal behavior of the experimenter; it has no effect upon the behavior of the experimental subject” (p. 338).

The hypothetical questions “What would deaf children be like if they had deaf instead of hearing parents?” or “What would deaf parents be like if they had the same average socioeconomic status as hearing parents?” are like those raised implicitly by the application of randomization tests to intact groups, as in Fisher’s (1936) test for the difference in height between Englishmen and Frenchmen (Chap. 7). In all these cases the futile hypotheticals result from the perceived need to achieve control in the outside world by adding or subtracting isolated characteristics of intact organisms. A question like “What would you be like if you were the opposite sex?” which would be implicit in a randomization test for sex differences on some measure, will have but ambiguous scientific import at best. Such questions, on the other hand, are often extremely useful clinically—which is to say, not that they lack epistemic relevance, but that what they point to is something more about ourselves than about the outside world (cf. Devereux, 1967). Like the techniques of standard projective tests, perhaps, they are subjectively revealing in proportion to the ambiguity of their objective reference. But the clinical import is not the end of the line, either: Through the invitation to consider other perspectives, we may well be prompted to change ourselves, an act of accommodation that, however modest, represents a greater advance in the understanding of sex differences than the results of a randomization test.¹⁴

With our assimilative focus on control, we have recognized only one half of the circle, and have failed to appreciate the paradoxical implications of pressing only in that direction. The paradoxes are dissolved if we relax our control enough to allow awareness to flow freely in both directions. In becoming more aware of ourselves, we remove the paradoxes of observer exclusion attending a more purely assimilative approach to knowing. And in accommodating to the diversity of nature, we obviate the paradoxes of attempts to force the equalization of inherently unequal individuals. The dilemmas of equalizing intact groups or individuals, through

¹⁴Pirsig’s (1974) concept of Quality as the cutting edge of consciousness makes a similar plea for accommodation to the object. The eloquence and power of his exposition derives partly from his provocative application of the concept to motorcycle maintenance. If the idea of making ourselves one with a machine is daunting to some, sensitive attunement to another person is no less delicate a task for others. Whichever comes more easily will depend on whether we grew up more like a person or a motorcycle.

matching or regression analysis, may be regarded as a special instance of the general problem confronting cross-cultural research. And it may be useful to consider further that the smallest minority “culture” is the individual. Our rather long experience in trying to develop culture-free measures of intelligence, of understanding of syntax, of self-esteem, or anything else, is by now enough to suggest that we are reaching the limits, at least for the time being, in these various attempts to reduce Other to Same. Levinas (1974), in fact, sees the other as requiring infinite intellectual accommodation, as constantly surpassing and confuting our concepts and systems. In cognitive terms, welcoming the Other as different is an act of self-questioning, in other words, an invitation to growth. By the principle of duality, when we compress and freeze the world into our categories, we freeze ourselves at the same time; so in fact, we are chronically dependent on the creativity (for Piaget; infinity, for Levinas) of the Other to keep us alive and growing.

The particular kind of knowers we make ourselves thus has implications for the kind of persons we become. Essentially, orthodox methodology is designed for use *by* as well as *on* black boxes. Taking objectivity to entail a denial of the subject, it assumes ideally no greater sensitivity on our part than that required to read a meter. Its epistemological egalitarianism aims to obscure differences in individual knowers, promoting the illusion of observer exclusion. In place of the connectedness to another that Benjamin (1988), Sampson (1985), and others have called for, it offers us a reliability coefficient. Ultimately its ominous image of the knower is of an insensitive brute in total control.

11.3.3 Objectivity, Skeuomorphosis, and the Problem of Scale

In Chap. 2 I discussed briefly, following Porter (1992a, 1992b), how the emergence of the concept of objectivity could be understood historically as the development of a technology of distance in response to the tremendous increase in scientific activity and communication in the seventeenth century. But this perspective raises the question of whether, or in what sense, objectivity may be intrinsic to the conduct of science as anything beyond a cottage industry.

There are notable precedents for such a conclusion in other fields. The chief theorist among those who have blamed bigness *per se* for the dependable destructiveness of large-scale operations is Leopold Kohr (1957/1978), though his student and popularizer E. F. Schumacher is much better known. The esthetic of smallness is not hard to appreciate. The small Midwestern towns, where everybody knows everybody else, children play unattended, and doors are left unlocked at night, for example, contrast appealingly with the big coastal cities, where many of us never know our nearest neighbors—unless one is growing up gay in one of those towns. Smallness by no means guarantees intersubjectivity; in fact, social homogeneity can thwart it as effectively as an overaccommodating parent. Small communities can, and commonly do, deal with difference by denying, destroying, or banishing it. I

would thus maintain, with Benjamin and against Kohr, that the problem is not so much bigness per se as the failure of intersubjectivity.

Indeed, to reverse all the troublesome transformations of scale might well unzip human evolution all the way back to the Stone Age. Jean Liedloff (1985) has given us an exquisite portrait of an intersubjective community if there ever was one—the Yequana Indians of southeastern Venezuela. It is hard to escape the inference, however, that it is their perfect contentment which leaves them no motivation for moving beyond the Stone Age. Disorders of domination and control attending the breakdown of intersubjectivity may not be the only impetus to exploration and expansion, but they are a strong one. The persistence of the Yequana way of life can also be attributed to their isolation. However much we romanticize such communities and exhort them to resist the blandishments of the West, their enthusiasm for electronics, rock music, brand-name jeans, and nuclear weapons evidently knows no bounds. (Some would say the coca-colonization of “developing” countries is paralleled by the californication of the “developing” states.) Kohr himself was well aware of the attractions of size, and realistic about the prospects for “the breakdown of nations.” His 10th chapter is entitled, “The Elimination of Great Powers: Can It Be Done?” The 11th, titled, “But Will It Be Done?” consists of the single word “No.”

In assessing the hazards of bigness, it is useful to sort them into internal and external. (a) Entities, especially political organizations, tend to become aggressive bullies just as soon as they get big enough to get away with it. Indeed, it is depressingly common for formerly oppressed groups, once they have gained power, to become oppressors themselves. In the natural world runaway populations sooner or later encounter constraints—viruses, for example, must be careful not to become too virulent, lest they destroy all their potential hosts—though these natural corrections are slower and clumsier than we might wish. With respect to human organizations, particularly nations, it is not clear what could ever check their growth at any predefined point except a still more powerful organization: just what we were trying to avoid. (b) The larger a group, the more difficult it is to satisfy all its members. For this reason unification, commonly urged as the path to peace, often leads to war instead. The dilemmas of centralized government schooling in this country provide a familiar example, where a single decision must suffice for everyone: whether to teach sex, creationism, and so on. There is no solution but to break up such a monolithic system and allow families to choose the kind of education they want. But at least this recognition gives us some control over the problem of bigness: We can be more careful than we have been about creating larger entities.

Robert Nozick (1974) perhaps expresses this point best:

Wittgenstein, Elizabeth Taylor, Bertrand Russell, Thomas Merton, Yogi Berra, Allen Ginsburg, Harry Wolfson, Thoreau, Casey Stengel, the Lubavitcher Rebbe, Picasso, Moses, Einstein, Hugh Hefner, Socrates, Henry Ford, Lenny Bruce, Baba Ram Dass, Gandhi, Sir Edmund Hillary, Raymond Lubitz, Buddha, Frank Sinatra, Columbus, Freud, Norman Mailer, Ayn Rand, Baron Rothschild, Ted Williams, Thomas Edison, H. L. Mencken, Thomas Jefferson, Ralph Ellison, Bobby Fischer, Emma Goldman, Peter Kropotkin, you, and your parents. Is there really *one* kind of life which is best for each of these people? (p. 310)

His own utopian proposal is precisely for a framework for utopias: namely, a structure that, being very loose and libertarian in its own right, supported fragmentation and diversity (including authoritarianism) among its components. With respect to science, Nozick's framework is paralleled by Koch's (1976) concept of communities of discourse, the point of which is the futility of trying to impose a uniform language or system of thought on a domain as unruly as, say, psychology. The degree to which such a concept is *not* presently realized in psychology is increasingly attributable to the extreme dominance of funding by a single agency, namely the federal government, which has served largely to sustain one paradigm and protect it from competing ideas.

Diederich (1991) more recently proposes what we could take as a friendly amendment to Nozick:

There is no *guarantee* that no group will be overrun by another group. But to call for a police in order to prevent this would be to give up libertarianism from the start. What anarchism thus suggests is that a free society should be built up *from the bottom*—by “association,” that is—not within a pre-regulated frame. (p. 223)

On a global level, of course, such an arrangement is what we have always had; but the idea, extended to lower levels, would incorporate much of Kohr's vision. As soon as we apprehend that communication between such entities is vital, however, we confront a paradox. Clarence Tripp (1975), illustrating his principle of resistance in sexual attraction, movingly describes the need of protozoa for sexual conjugation with others sufficiently different from (resistant to) themselves in order to be refreshed for several generations of asexual reproduction by simple fission. Communities, if they remain insular, similarly risk becoming stale and too homogeneous for effective intersubjectivity. Unrestricted communication between them, however, would tend to imply that their boundaries were artificial, even pointless, so that they were really all just one big community after all. The definition of a political entity and its size will thus be subject to fluctuation and ambiguity. Iris Young (1990), in fact, criticizes the ideal of face-to-face community as detemporalizing, as well as totalizing and unrealistic; she also notes dryly how attached many of its advocates are to living in Boston or San Francisco. If bigness, in short, has led to impersonality, the price of staying small, in these contexts, is stagnation. That is surely no less true of science than of society.

To conclude, however, that we confront the choice of nothing but two evils would be unduly pessimistic. It goes without saying that, just as big-city life holds its attractions even for “small is beautiful” advocates, so there are benefits of large-scale science. What is relevant in the present context is that the same processes which give rise to objectivity also make possible intersubjectivity.

In general, communication—structural coupling, to use Maturana and Varela's (1980) generic terms—requires the articulation of a boundary, the surface, literal or metaphorical, at which organisms interact. Autopoiesis defines the first such boundary; consciousness, the second, creating an inner and outer world. The outer world will henceforth be experienced as real and concrete. Language we may count as a third such development toward externalization. The reduction of experience to language is so habitual and automatic that we tend to notice it only when it breaks

down, when “it doesn’t translate.” We judge whether we are in touch with our experience according to whether we have found the right words. And “the world” is now the world as apprehended linguistically.

These externalizing developments were not necessarily instigated by transformations of scale, and their skeuomorphosis is relatively complete. In changes beyond this point, however, the link between skeuomorphosis and transformations of scale, described in the more limited context of Chap. 2, becomes apparent, as does the “codependent arising” of subject and object. The examples of personal names and clothing were mentioned in that chapter. Within language, the gradual development of elaborated code (Bernstein, 1971), perhaps occurring first in ancient Greece, requires little additional comment as an illustration of the principle that any instrumentality that subserves interaction over greater physical or personal distance potentially furthers both objectivity and subjectivity; and what the price system did to our concept of value was already discussed in Chap. 2.

It remains to revisit the concept of scientific objectivity in this light. We have already noted, in fact, that the dramatic expansion of scientific activity in the sixteenth and seventeenth centuries had the effect of increasing subjectivity as well as objectivity, with the awesome revelations through the telescope and the microscope reinforcing the displacing effects of the recent geographical discoveries. The disquietude attending enhanced consciousness of oneself as subject could in principle be assuaged by enhanced intersubjectivity, a greater capacity for encountering others in their uniqueness, but transformations of scale propel us toward objectivity instead. Interaction at a distance pulls strongly toward the surface; the interior, the subtle, the idiosyncratic yield to the overt, the quantifiable, and the common as the basis for transactions. Objectivity, rooted in what is common, enjoys a further advantage over intersubjectivity, rooted in differentness, just because sameness is easier for us to deal with than differentness. (Notice how often proposals for conflict resolution entail that we all be alike.) As the ever-ready stand-in for intersubjectivity, objectivity thus functions like carbon monoxide, with its greater affinity than oxygen for hemoglobin. The effects of the substitution are subtle enough that we don’t notice them until it’s too late. As a side consequence, we are left with two concepts of objectivity, which are not always clearly distinguished, even as one is a term of approbation and the other of abuse. In the former, more fundamental, sense, it refers to the capacity for taking a perspective on oneself, for seeing oneself as an object; in the latter, it refers to seeing oneself and others as *mere* objects.

Gebser (1985/1949–1953, p. 355) argues that “something can be superseded and integrated only when concretized.” The problem is just that there is no guarantee that supersession and integration will be the outcomes; we may simply go on operating on concretes.

11.3.4 *The Scarecrows¹⁵ of Relativism and Anarchy*

It is probably obvious by now what these poor creatures are in for.

A major contribution of the theories of Benjamin and Dinnerstein is the deconstruction of various dichotomies pervading our culture, revealing their respective poles as straw persons. Both anarchy and relativism have been widely regarded as unthinkable, even as the logic of postmodernism pointed directly toward them. We reject Feyerabend's formulation "Anything goes" as implying that nothing matters; it makes no difference what we do in scientific work. I am not sure Feyerabend anticipated such a reading; in any event he later (1987) added what should have been unnecessary, that the principle "Anything goes" does not make scientific work *easier*. As Jeffreys (1939/1961) and others have pointed out, it is usually hard enough to formulate even one theory to fit the data. "Anything goes," as Wartofsky (1991) put it, in a recent festschrift for Feyerabend (Munévar, 1991), means "not that we don't care, but are willing to try anything and see how it goes" (p. 28).

The reason, of course, why anarchy should have been so unthinkable is just that the breakdown of intersubjective tension has obliterated any legitimate concept of self-interest or self-responsibility, either cognitive or personal, so that we can no longer conceive of alternatives to strong external control, on the one hand, or the familiar caricature of "ruthless, unbridled selfishness," on the other. The implication—still widely invoked as a justification for all sorts of policing procedures, including statistical inference—is that, unchecked by external controls, we would all, as persons or as scientists, run wild, abusing our neighbors or our data. The chronicity of our abdication, of course, has important implications for the process of change: The withdrawal of those external controls, epistemological or political, will not be feasible except in proportion as they are replaced by natural bonds to reality and to other people. Without the latter development, anarchy in either realm would indeed be chaos and destruction, as is usually feared.

A recent study by Alifano (1995) illustrates how these ties to the world and other people entail a high level of responsibility, sensitivity, and sophistication in judgment. The therapists who participated in his research agreed that empathic surrender by the therapist was under the right conditions a very powerful and rewarding move; but it is also a high-risk practice, in directly violating well-established rules about maintenance of boundaries between client and therapist. The consensus was accordingly that its use should be restricted to experienced practitioners (a new rule presumably to be challenged in its own time). In other fields as well, one can see the evolution, for similar reasons, of "secrets of the tantra," not to be shared with novices.

It is also the breakdown of intersubjective tension which underlies the fear of relativism. The usual formulation—"If there are no absolutes, then Hitler is the moral equivalent of Gandhi"—forces knowledge into a dichotomy of certainty versus nihilism. But the idea that there is nothing to hold us on the "slippery slope" into

¹⁵I am indebted to Jürgen Kremer (1993) for this apt term.

the abyss amounts to the claim that nothing matters; the world is empty. The equation of relativism with nihilism, in insisting that nothing less than proof is adequate, functions as the epistemological equivalent of the minimum-wage law: Just as that law has the effect of knocking the bottom rungs out of the economic ladder, making it impossible to get started unless one has already enjoyed some measure of success, so the demand for proof as the first step makes it impossible ever to get off the ground in acquiring knowledge. Things matter—to human beings, and to living control systems in general. Making and adjudicating claims about the world, including ourselves, need not await an ultimate understanding of biology.

The loss of absolutes seems awful mainly if we grew up expecting something else, namely that grown-ups knew all the answers. Indeed, so effective is our message to children that there are rules and authorities for everything, much of the process of growing up is a matter of getting over the shock of discovering otherwise. The point is nevertheless a tricky one, for today's intellectual climate occasionally presents us with the spectacle of a glib or flaccid relativism that blandly declares everything to be a matter of mere opinion, and therefore denies the possibility of any real conflict or criticism. As a breakdown of intersubjective tension in the opposite direction, this phenomenon usefully reminds us just how delicate a task the maintenance of that tension is.

In a much more sophisticated response, postmodernism has tended to embrace pluralism in the alluring metaphor of the dance, with transcendence to be achieved by skipping about nimbly between perspectives. Yet, as Susan Bordo (1990) and Don Johnson (1994) argue, this attempt to attain the “view from everywhere” cannot succeed, for it denies our embodiment, our embeddedness in history as well as a particular corpus, which limits the moves we can make. Sampson (1981) may be right that putting ourselves in another's shoes in our imagination is no substitute for actually putting on the shoes and walking around in them, and that it is arrogant to suppose otherwise; yet intersubjective recognition may be as close as we can come.

In general, where the failure of intersubjectivity has left us with false choices, the task confronting us is the recognition, in both the cognitive and diplomatic sense, of the disowned aspect, and the establishment of an equilibration, a dialogue, between different aspects of the self, between self and world, and between one another. In personal terms, however, the task will in many cases be heroic. In the first place, learning, as a species of self-change, is threatening, in proportion to our investment in the current order. Secondly, in any polar situation where estrangement and suppression have proceeded so far, it becomes very difficult to grant any legitimacy to the rejected pole without feeling that we are opening the door to a flood—in this case of denied femininity and vulnerability. The potential benefit for us, to the extent we can stand it, is an unaccustomed personal integrity, in the root meaning of the term, and cognitive enfranchisement. We cannot really know the emotional, or the intellectual, price we have paid until that healing is begun.

11.4 Biomedical Research Without a Biomedical Model

Assuming all these obstacles transcended (or simply passed over without being resolved), what would a genuine science of psychological and medical research look like? It is essential to realize that practically all of our research heretofore has been based on models that were not biomedical, but epidemiological. If we want to know, for example, whether a new statin lowers cholesterol, the statistical model is

$$y_{ij} = \mu + bx_j + e_{ij}, i = 1, \dots, n_j, j = 1, 2.$$

Here, j indexes the statistical treatment (yes or no), and i indexes individuals within group; x is the statin dose, b is the coefficient (presumed negative) to be estimated, y is the measured cholesterol; μ is the population mean cholesterol; and e is “error,” the individual deviations in response from the average. But the drug doesn’t act on a random aggregate; it acts on different individuals, with different effects. The model estimates only the average of a random aggregate; *it is not a biological model*. It is true that medicine in recent decades has shifted, under the influence of the prevailing methodology, to a public health, epidemiological, perspective; cholesterol, smoking, and sexual diseases are treated as epidemics; but such treatment does not make medicine into epidemiology, and does little to advance understanding. Michel Accad (2017), in his book *Moving Mountains*, explores the implications of treating populations rather than patients.

The focus on population parameters has truly bizarre implications in research practice. Individuals exist only as abstractions, not as biological beings. On the modern seventeenth-century model of knowledge, they are conceived as aggregates of characteristics which can be added or subtracted at will, of attributes which are as detachable as flags from steamboats (Chaps. 2 and 4). The simplest illustration is the treatment of outliers.

Suppose your doctor said, “Your blood pressure is 228. That’s completely outside the range of what we commonly see, so I’m just going to record that as 150.” I suspect that would be considered poor medical practice. The funny thing is that it’s considered good *scientific* practice. In medical practice we often do change data, to make sure that someone qualifies for disability payments, or meets some other arbitrary bureaucratic cut-off we want them to meet. But you would think that in science, of all places, we wouldn’t have any need to change the data. In a major study in which I participated, involving a sample of about 180 HIV+ men, most cortisol measurements were less than 4, and many were less than 1. But there were a few values ranging up into the 100s. There were several men whose measurements were consistently in the 100s, any time of day, over a span of many months; so they were clearly not measurement or recording errors. We changed them all to 100 (and said so). One interesting consequence of this practice is that other investigators, when they get values in the 100s, think, “Nobody has ever reported values like this before. They must be a mistake; we’d better change them, and keep it all hushed up.” A totally false picture of human variation thus gets built up very quickly. And that

false picture is widely taken to be reality. Back in the 1950s, Roger J. Williams, who won the Nobel Prize for the discovery of Vitamin B5, wrote a book called *Biochemical Individuality* (Williams, 1956/1998). He found that, whatever biological parameter he examined, from cholesterol to spleen size, there was variation of an order of magnitude or more in very small samples of 10 or 20 individuals. But his work seems never to have attracted any attention. Nature appears to us as uniform because we never allow it to be anything else.

The biological point is the assumption that we can change values like these at will, without affecting any others. But there is a real question whether these data points as a vector represented biologically viable beings. The absurdities discussed above, with the Meadow (1967/1968) matching study or the use of covariate adjustment on intact groups, are further consequences of population models; they would not arise in a study of individuals.

Research workers in psychology and medicine imagine themselves to be playing scientist when they do their regressions, just like the big people in physics and chemistry; the difference is that the latter are paying attention to the meaning of the numbers and their manipulations. A pair of example makes the point.

Mailen, Smith, and Ferris (1971) measured the relation between temperature and the solubility of plutonium fluoride in a lithium-beryllium fluoride solvent. Physical theory predicts a linear relation, across the range of temperatures studied, between 1000°C and $-\log_{10}$ solubility. This example is interesting, because the errors are too big to be errors of measurement. The deviations from linearity probably reflect mostly variations in time the solution was allowed to reach equilibrium; and it is a reasonable presumption that adding that variable to the equation would remove the excess error and make the relationship linear in three dimensions—a presumption which could be checked by physical experiment (Fig. 11.1).

The following graph, of the relation between number of physical symptoms and depression in a sample of 296 women, shows what psychologists do. The correlation, 0.52, is rather high for psychological research. But the graph still exhibits almost nothing but “error”; 27% seems a generous estimate of amount of variance in depression “accounted for” by symptoms. A linear relationship is in no way an adequate summary of what is going on here. But the more important point is that there is no reason to assume that, if we merely added enough variables to the equation, the true linear relationship between symptoms and depression would be revealed. There’s no theory which holds that the relation between symptoms and depression ought to be linear across any given sample of *different people* (Fig. 11.2).

Aschbacher et al. (2012) is an example of a recent study looking at parameters within an individual. These authors were concerned to predict cortisol levels at 10-minute intervals over a 24-hour period. They were able to predict cortisol values with impressive accuracy for a very simple model using only the previous cortisol measurement and the level of ACTH:

$$\frac{dC}{dt}(t) = \lambda_i - \lambda_c(t) + \lambda_A A(t)$$

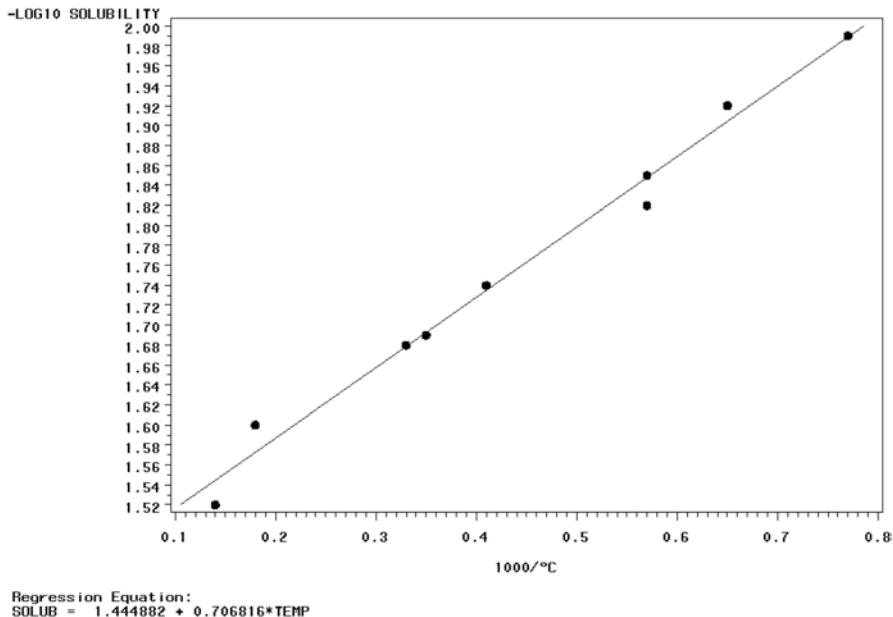


Fig. 11.1 Solubility of plutonium fluoride as a function of temperature

Here $\frac{dC}{dt}(t)$ is the rate of change of cortisol, $A(t)$ is the amount of ACTH secreted at time t , λ_A is a sensitivity coefficient for the adrenal cortex to the ACTH signal, $\lambda_C(t)$ is the cortisol clearance rates and sensitivity of cortisol receptors, and λ_i is “everything else” that could affect cortisol. These authors used “model-based predictive control” (MPC; Ben-Zvi, Vernon, & Broderick, 2009; Gupta, Askakson, Gurbaxani, & Vernon, 2007; Savić & Jelić, 2005), based on “generalized predictive control” (GPC; Clarke, Mohtadi, & Tuffs, 1987). But the prediction is from the point of view of an observer of the system. The hypothalamic-pituitary system doesn’t do any prediction. Nor is it solving differential equations in real time—although, in fairness, water molecules entering a pipe of decreasing diameter don’t need to solve a differential equation to know where to go; physical theory specifies the parameters for the equation describing their motion. Theory in the cortisol case is thin enough that it specifies only that ACTH acts to increase it and previous cortisol levels act to decrease it, both at rates to be determined empirically. But if the simulation amounts to a kind of “clever Hans” trick, it nevertheless represents a substantial advance over previous research, based on modeling aggregates of individuals.

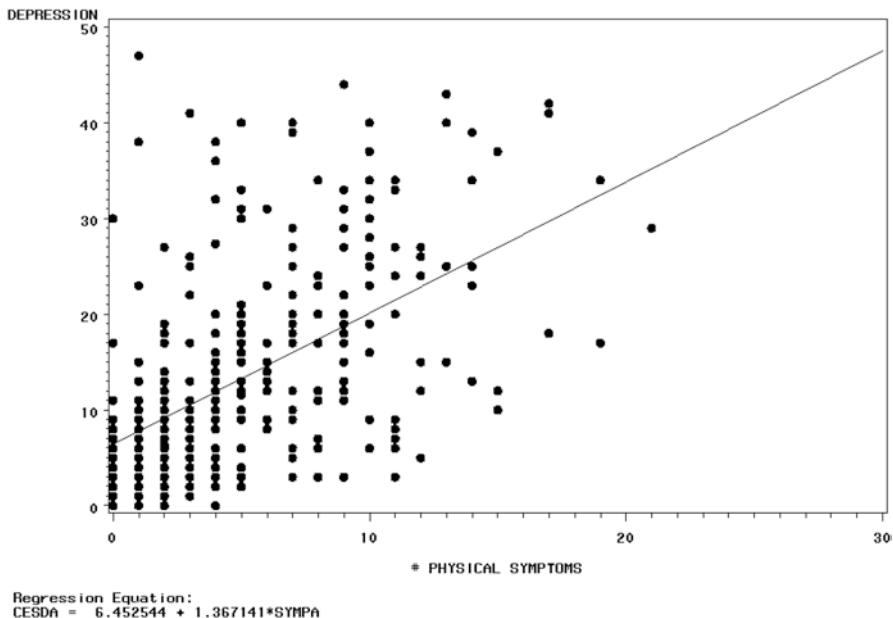


Fig. 11.2 Depression as a function of number of physical symptoms in 296 maternal caregivers

11.5 Postscript on Perceptual Control Theory

I have encountered only one general theory which is even remotely adequate to account for biological and psychological functioning; that is the Perceptual Control Theory (PCT) of Bill Powers (1973, 1978). It is radical enough to require a brief introduction, which I adapt from a previous article.

The traditional view—both folk and scientific (whether behaviorist, psychoanalytic, or whatever)—is that people control (or try to control) their actions. PCT holds that what we control are our *perceptions*, using that term broadly enough to encompass everything from the nonconscious control of variables like body temperature or blood pressure to constructs like self-esteem.

The stock illustration is driving. When we are driving, we are not, as the naïve might assume, controlling the position of the steering wheel. We can't plan where we want the steering wheel to be even a second in advance. What we actually control is more like the picture in the windshield—keeping in the center of the lane, maintaining a certain distance from the car in front, and so on. Our driving behaviors consist of whatever is necessary to maintain the desired perception. The desired perception constitutes a reference signal, against which we compare our existing perception. If I perceive the car drifting to the left of where I want to be, I act to move it to the right. Inexperienced drivers may overcorrect, and require a cycle or two to stabilize the position. But for anyone, at any stage of the process, discrepancies, in a closed feedback loop, lead to outputs—behaviors—in the opposite direction to reduce the error, and so on and on around the loop. What we're trying to control is not output but input (perception)...

Powers suggests ten or eleven levels of perception control in humans. Higher-order control systems set reference levels for lower ones. As I write this in longhand, low-level

control systems are maintaining a certain pressure between my fingers and the pen, while others control the formation of letters in the sequence I intend. All this in the service, higher up the hierarchy, of goals like advancing ... awareness of PCT. At the same time, other control systems are keeping me rocking in the porch swing and maintaining my balance and my blood sugar levels.

At the base of the hierarchy are biological variables, like blood sugar and body temperature, that have to be maintained within a certain range for survival. Deviations from the reference state of these variables constitute *intrinsic error*, which we will act, at one level or another, to remove. If we don't know what to do—if none of the systems we have developed for controlling a variable is working—we start trying things at random. If we stumble on something that works, we have a new control system. The same mechanism—reorganization—comes into play in novel situations, and is involved in learning. A simple example of Powers': If we are approaching a door and don't know whether to push or pull, we simply try one or the other. A simple and elegant demonstration of reorganization is provided by the single-cell *E. coli*, which has but two means of locomotion. It can go straight ahead by moving its flagellae together, or it can tumble by moving them asynchronously. These capabilities are sufficient to propel it, with about 70% efficiency, along an increasing sugar gradient. So long as the environment is getting sweeter, it keeps going; when the sweetness decreases, it tumbles and takes off in a new direction, at random.

"Control of perception," of course, does not mean that I can willfully see an apple as an orange (without simply substituting an orange for the apple of my eye). The states of variables are controlled, to the extent anything is. For example, when I move an apple to my mouth, what I control is perception of its position. Many perceptions cannot be controlled, however I might wish otherwise. That is our principal evidence for an external world. It's all perception, but it's not all of my making.

The reference signal, setting the goal for any particular control loop, embodies the concept of purpose. One of the significant achievements of control theory is thus the clear reconciliation of mechanism and purpose. Purpose has hitherto been generally excluded from the life sciences, as inconsistent with mechanism, but the problem has been simply that our concept of mechanism was too limited.

The closed negative-feedback control loop also embodies circular rather than linear causation. Perceptual inputs and behavioral outputs affect each other, with near simultaneity at the lower levels. Traditional theories, like behaviorism, cut the causal loop in such a way as to make it appear that input (environmental stimuli) causes output (behavior)—although our actions also affect what we subsequently perceive. In economics, for example, we commonly speak causally of incentives as if they determine behavior, but it is only the fact that people typically have indefinitely high reference level for money that makes environmental conditions incentives or not. Although the controlled variable is obvious in the case, it can often be difficult to discern what variable people are actually controlling. It may take a skilled psychotherapist to find out...

Control theory per se is not new; it was developed by engineers around 90 years ago. There have been a number of thinkers since then who have glimpsed its relevance for biology and psychology, but in many cases, they didn't understand control theory well enough to make real use of it. Powers is the first to have worked out a systematic model of life, and of human functioning, specifically. It has been tremendously fashionable in the last 50 years for psychologists to propose "models" consisting of boxes linked with arrows, but when Powers, an engineer, presents a model, he's talking about something that you can build and that will work in the specified way. And his model has remained consistent with everything that has been learned about the neural, endocrine, and other systems of the body in the last 40 years.

That's much more than any rival theory can claim. In 1943, at the start of the digital revolution, McCulloch and Pitts (1943) proposed that the nervous system might operate as a digital computer, with the firing or not firing of neurons constituting the binary basis of

digital arithmetic and logical circuits. Psychologists, except for Powers, have never looked back on that initial assumption. One consequence is that models of human functioning have focused overwhelmingly on high-level cognitive operations, like chess playing, at which digital computers excel. In 1964, my introductory psychology professor observed that we still couldn't explain how a rat scratches itself. His observation still holds for mainstream psychology 50 years later. We're nowhere in terms of being able to model simple animal—or insect—behavior such as walking over uneven terrain. Current models of actions as simple as reaching out to pick up a glass of water require calculating the desired trajectory and thus the inverse kinematics, entailing the solution of very large systems of nonlinear differential equations. None of us can begin to do that, especially in real time—yet we assume that the nervous system of a rat or a dragonfly can. Powers, observing that neurons fire at (more or less) continuously varying rates, argues that the nervous system is an *analog* computer. The neural architecture for such operations, using negative-feedback control systems, becomes extremely simple by comparison—within the capability of an ant...

It is the essence of a control system that when something disturbs a controlled variable, the system acts to correct the disturbance, to reduce the error; it pushes back. I can boost my thyroid level temporarily by taking thyroid capsules, but my thyroid gland, perceiving a higher level than its current reference level, will shut down production. (Not being a yogi, I don't have conscious access to my thyroid level.) If someone says something that threatens my self-esteem, it will be hard for me not to try to correct the perceived error, the departure from my reference level. Similarly, if I myself do something that disturbs my preferred perception of self-esteem, I will rationalize that behavior or in other ways act to reduce the perceived error.

Conflict occurs when two control systems attempt to control the same variable at different reference levels. That's a problem within, as well as between, persons. If reference levels of two systems are far apart, relative to error sensitivity, both systems will output their maxima, canceling each other and leaving neither in control. If a disturbance moves the controlled quantity toward the reference for one system, the system will relax, so the quantity gets pulled back in the opposite direction. Powers offers the example of a man who has the goals both of being “assertive” and of being “nice.” If he speaks up for himself, for instance in asking for a raise, the error created by his reference level for niceness will lead him to undo that act in some way, like a smile suggesting that he really didn't mean it. Then he will rebuke himself for being wishy-washy, and so on in endless vacillation.

We have been taught to deal with internal conflict by self-control, overcoming particular desires or fears by force of will. But this approach simply pits one control system against another. Because this approach is arbitrary, in the sense that it takes no account of the goals the behavior is helping to control, it will typically induce further conflicts elsewhere in the system. Resolution usually entails moving to a higher level, from which the system may be surveyed; Powers [developed] a technique he calls the Method of Levels [Carey, 2006; Mansell & Goldstein, 2020] to assist in that process (Acree, 2002, pp. 35–37).

So-called empirical research on PCT is still in its infancy. Bruce Nevin (2020) has been doing good work studying language from a PCT point of view, and Frans Plooij (2020) has applied PCT thoughtfully to the study of development. A more concrete example is Henry Yin's (2014) work in neuroscience, demonstrating that the velocity of movement is controlled in the basal ganglia. It is important to note that Yin's model could easily be implemented by the nervous system, since the requisite computations can be done with only a handful of neurons, if the nervous system is operating as an analog computer.

PCT has also been limited, at the physiological level, largely to the study of motor movement. But there are other systems that could benefit from a similar analysis. The hypothalamic-pituitary system and the immune system, for example, are

analogously organized as systems for the control of certain quantities. Analyses under the prevailing linear paradigm, pursuing questions like whether more cortisol is a good thing or a bad thing, are otiose; the relevant question is always whether the amount is optimal for a given individual system under given specific conditions. Generalizations can profitably be made along these lines, but hardly across individuals and conditions. There are no useful insights to be gained by averaging across individuals or conditions. If we put medicine—and psychology—back on a footing of understanding biological systems, we will have no more need of statistical inference in these fields.

References

- Accad, M. (2017). *Moving mountains: A Socratic challenge to the theory and practice of population medicine*. College Station, TX: Green Publishing House.
- Acree, M. (2002, May). Perception, control and anarchy. *Liberty*, 16(5), 35–38, 40.
- Alifano, R. (1995). *Empathic surrender: A phenomenological investigation*. Unpublished doctoral dissertation, California Institute of Integral Studies.
- Ancillon (J. P. F.). (1794). Doutes sur les bases du calcul des probabilités [Doubts on the bases of the calculus of probabilities]. *Mémoires de l'Académie Royale des Sciences et Belles-lettres de Berlin*, 3, 3–32.
- Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In. Universities-National Bureau Committee for Economic Research. In Committee o Economic Growth of the Social Science Research Council (Ed.), *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton, NJ: Princeton University Press.
- Aschbacher, K., Adam, E. K., Crofford, L. J., Kemeny, M. E., Demittrack, M. A., & Ben-Zvi, A. (2012). Linking disease symptoms and subtypes with personalized systems-based phenotypes: A proof of concept study. *Brain, Behavior, and Immunity*, 26, 1047–1056.
- Benjamin, J. (1988). *The bonds of love: Psychoanalysis, feminism, and the problem of domination*. New York, NY: Pantheon.
- Ben-Zvi, A., Vernon, S. D., & Broderick, G. (2009). Model-based therapeutic correction of hypothalamic-pituitary-adrenal axis dysfunction. *PLoS Computational Biology*, 5(1), e1000273. <https://doi.org/10.1371/journal.pcbi.1000273>
- Berman, M. (1989). *Coming to our senses: Body and spirit in the hidden history of the West*. New York, NY: Simon and Schuster.
- Bernoulli, J. (1713). *Ars conjectandi [The art of conjecturing]*. Basel, Switzerland: Thurneysen.
- Bernstein, B. (1971). *Class, codes, and control*. London, UK: Routledge, Kegan, Paul.
- Bordo, S. (1990). Feminism, postmodernism, and gender-scepticism. In L. J. Nicholson (Ed.), *Feminism/postmodernism* (pp. 133–156). New York, NY: Routledge.
- Bordo, S. R. (1987). *The flight to objectivity: Essays on Cartesianism and culture*. Albany, NY: State University of New York Press.
- Branden, B. (1962). *The principles of efficient thinking*. New York, NY: Nathaniel Branden Institute.
- Branden, N. (1969). *The psychology of self-esteem*. Los Angeles, CA: Nash Publishing.
- Branden, N. (1984). *Honoring the self: Personal integrity and the heroic potentials of human nature*. Los Angeles, CA: J. P. Tarcher.
- Byock, J. (2001). *Viking age Iceland*. New York, NY: Penguin.
- Callahan, R. E. (1962). *Education and the cult of efficiency*. Chicago, IL: University of Chicago Press.

- Carey, T. A. (2006). *The method of levels: How to do psychotherapy without getting in the way*. Hayward, CA: Living Control Systems Publishing.
- Carson, K. A. (2008). *Organization theory: A libertarian perspective*. Booksurge.
- Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, 103, 105–110.
- Clarke, D. W., Mohtadi, C., & Tuffs, P. S. (1987). Generalized predictive control—Part I. The basic algorithm. *Automatica*, 23, 137–148.
- Code, L. (1992). Who cares? The poverty of objectivism for a moral epistemology. *Annals of Scholarship*, 9, 1–17.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65, 145–153.
- Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, 33, 107–112.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York, NY: Academic Press.
- Cohen, L. J. (1977). *The probable and the provable*. Oxford, UK: Clarendon Press.
- Comte, A. (1864). *Cours de philosophie positive* [Course in positive philosophy] (2nd ed.). Paris, France: Baillière. (1st ed. 1830–1842).
- Danziger, K. (1991). *Constructing the subject: Historical origins of psychological research*. Cambridge, UK: Cambridge University Press.
- Denzin, N. K., & Lincoln, Y. S. (Eds.). (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Devereux, G. (1967). *From anxiety to method in the behavioural sciences*. The Hague, The Netherlands: Mouton.
- Diamond, I., & Quinby, L. (Eds.). (1988). *Feminism and Foucault: Reflections on resistance*. Boston, MA: Northeastern University Press.
- Diederich, W. (1991). Obituary on the “anarchist” Paul Feyerabend. In G. Munévar (Ed.), *Beyond reason: Essays on the philosophy of Paul Feyerabend* (pp. 213–224). Dordrecht, The Netherlands: Kluwer Academic.
- Dinnerstein, D. (1976). *The mermaid and the minotaur: Sexual arrangements and human malaise*. New York, NY: Harper and Row.
- Dreyfus, H. L., & Dreyfus, S. E. (1992). What is moral maturity? Towards a phenomenology of ethical expertise. In J. Ogilvy (Ed.), *Revisioning philosophy* (pp. 111–131). Albany, NY: State University of New York Press.
- Fatalité* [Fatality]. (1756). In D. Diderot & J. le R. d'Alembert (Eds.), *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers* [Encyclopedia, or reasoned dictionary of sciences, arts, and crafts] (Vol. 6, pp. 422–429). Paris: Briasson, David, le Breton, et Durand.
- Fetzer, J. H. (1974). Statistical probabilities: Single-case propensities vs. long-run frequencies. In W. Leinfellner & E. Köhler (Eds.), *Developments in the methodology of social science* (pp. 387–397). Dordrecht, The Netherlands: Reidel.
- Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge*. London, UK: New Left Books.
- Feyerabend, P. K. (1987). *Farewell to reason*. London, UK: Verso.
- Fine, T. L. (1973). *Theories of probability*. New York, NY: Academic Press.
- Fisher, R. A. (1936). “The coefficient of racial likeness” and the future of craniometry. *Journal of the Royal Anthropological Institute*, 66, 57–63.
- Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society (Series B)*, 17, 69–78.
- Forward, J., Canter, R., & Kirsch, N. (1976). Role-enactment and deception methodologies. *American Psychologist*, 31, 595–604.
- Foucault, M. (1973). *The order of things*. New York, NY: Random House.
- Foucault, M. (1979). *Discipline and punish: The birth of the prison*. New York, NY: Random House.
- Friedman, D. (1989). *The machinery of freedom* (2nd ed.). La Salle, IL: Open Court.
- Gans, D. (1988, April). Ted Nelson and the ultimate information machine. *MicroTimes*, 53–57.

- Gebser, J. (1985). *The everpresent origin* (Trans. N. Barstad). Athens: Ohio University Press. (Original work published 1949-1953)
- Gergen, K. (1994). Exploring the postmodern: Perils or potentials? *American Psychologist*, 49, 412-416.
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1, 198-218.
- Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition*, 43, 127-171.
- Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.
- Giorgi, A. (Ed.). (1985). *Phenomenology and psychological research*. Pittsburgh, PA: Duquesne University Press.
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Chicago, IL: Aldine.
- Good, I. J. (1990). A quantal hypothesis for hadrons and the judging of physical numerology. In G. Grimmet & D. J. A. Welsh (Eds.), *Disorder in physical systems: Essays in honour of John M. Hammersley* (pp. 129-165). London, UK: Oxford University Press.
- Good, I. J. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, 87, 597-606.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24, 411-435.
- Gupta, S., Askakson, E., Gurbaxani, B. M., & Vernon, S. D. (2007). Inclusion of the glucocorticoid receptor in a hypothalamic pituitary adrenal axis model reveals bistability. *Theoretical Biology and Medical Modeling*, 4, 8. <https://doi.org/10.1186/1742-4682-4-8>
- Harding, S. (1986). *The science question in feminism*. Ithaca, NY: Cornell University Press.
- Harding, S. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Ithaca, NY: Cornell University Press.
- Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehart, Winston.
- Horrobin, D. F. (1982). Peer review: A philosophically faulty concept which is proving disastrous for science. *The Behavioral and Brain Sciences*, 5, 217-218.
- Husserl, E. (1989). Ideas pertaining to a pure phenomenology and to a phenomenological philosophy. Second book: *Studies in the phenomenology of constitution* (R. Rojcewicz & A. Schuwer, Trans.). Dordrecht, Netherlands: Kluwer Academic. (Original work published 1952).
- Ihde, D. (1977). *Experimental phenomenology: An introduction*. New York, NY: Capricorn Books.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59, 12-19.
- Jaggar, A. M., & Bordo, S. R. (Eds.). (1989). *Gender/body/knowledge: Feminist reconstructions of being and knowing*. New Brunswick, NJ: Rutgers University Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Clarendon Press. (1st ed., 1939).
- Johnson, D. H. (1994). *Body, spirit, and democracy*. Berkeley, CA: North Atlantic Books.
- Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.
- Kealey, T. (2021). Terence Kealey on the myths of public funding of science. *The Accad and Koka Report, Episode, 159*.
- Kempthorne, O. (1972). Theories of inference and data analysis. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor* (pp. 167-191). Ames, IA: Iowa State University Press.
- Kempthorne, O. (1976). Statistics and the philosophers. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Foundations and philosophy of statistical inference) (Vol. 2, pp. 273-314). Dordrecht, The Netherlands: Reidel.
- Kendall, M. G. (1949). Reconciliation of theories of probability. *Biometrika*, 36, 101-116.
- Keynes, J. M. (1973). *A treatise on probability*. New York, NY: St. Martins Press. (Original work published 1921).

- Keys, J. (1972). *Only two can play this game*. New York, NY: Julian Press.
- Koch, S. (1976). Language communities, search cells, and the psychological studies. *Nebraska Symposium on Motivation*, 23, 477–559.
- Koestler, A. (1964). *The act of creation*. New York, NY: Macmillan.
- Kohlberg, L. (1969). Stage and sequence: The cognitive-development approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research*. Chicago, IL: Rand McNally.
- Kohr, L. (1978). *The breakdown of nations*. New York, NY: Dutton. (Original work published 1957).
- Kremer, J. W. (1993). Evolution as world renewal: A shamanic concourse about dissociative Eurocentric cultural practices. Unpublished manuscript.
- Langer, S. K. (1967). *Mind: An essay on human feeling* (Vol. 1). Baltimore, MA: Johns Hopkins Press.
- Le Cam, L. (1977). A note on metastatistics, or “An essay toward stating a problem in the doctrine of chances.” *Synthèse*, 36, 133–160.
- Levi, E. H. (1949). *An introduction to legal reasoning*. Chicago, IL: University of Chicago Press.
- Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, 61, 378–382.
- Levinas, E. (1974). *Totalité et infini: Essai sur l'exteriorité [Totality and infinity: An essay on exteriority]* (4th ed.). The Hague, The Netherlands: Martinus Nijhoff.
- Liedloff, J. (1985). *The continuum concept* (rev. ed.). Reading, MA: Addison-Wesley.
- Lord, F. M. (1969). Statistical adjustments when comparing pre-existing groups. *Psychological Bulletin*, 72, 336–337.
- Lucas, J. R. (1970). *The concept of probability*. Oxford, UK: Clarendon Press.
- Mailen, J. C., Smith, F. J., & Ferris, L. M. (1971). Solubility of PuF₃ in molten 2 LiF-BeF₂. *Journal of Chemical and Engineering Data*, 16, 68–69.
- Mainland, D. (1960). The use and misuse of statistics in medical publications. *Clinical Pharmacology and Therapeutics*, 1, 412–422.
- Mansell, W., & Goldstein, D. M. (2020). Method of levels therapy. In W. Mansell (Ed.), *The interdisciplinary handbook of Perceptual Control Theory: Living control systems IV* (pp. 503–515). London, UK: Elsevier.
- Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht, The Netherlands: Reidel.
- Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge, UK: Cambridge University Press.
- McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133.
- Meadow, K. P. (1968). The effect of early manual communication and family climate on the deaf child's development (Doctoral dissertation, University of California, Berkeley, 1967). *Dissertation Abstracts International*, 28, 4295A–4296A. (University Microfilms No. 68-05785).
- Meadow, K. P. (1969). Self-image, family climate, and deafness. *Social Forces*, 47, 428–438.
- Medin, D. L., & Ross, B. H. (1992). *Cognitive psychology*. Fort Worth, TX: Harcourt Brace Jovanovich.
- Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology*, 64, 553–557.
- Morse, J. M. (1994). Designing funded qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 220–235). Thousand Oaks, CA: Sage.
- Munévar, G. (Ed.). (1991). *Beyond reason: Essays on the philosophy of Paul Feyerabend*. Dordrecht, The Netherlands: Kluwer Academic.
- Murray, C. (1988). *In pursuit of happiness and good government*. New York, NY: Simon and Schuster.
- Murray, L. W., & Dosser, J. D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology*, 34, 68–72.

- Nevin, B. (2020). Language and thought as control of perception. In W. Mansell (Ed.), *The interdisciplinary handbook of Perceptual Control Theory: Living control systems IV* (pp. 351–459). London, UK: Elsevier.
- Nicholson, L. J. (Ed.). (1990). *Feminism/postmodernism*. New York, NY: Routledge.
- Nozick, R. (1974). *Anarchy, state, and utopia*. New York, NY: Basic Books.
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.
- Orne, M. T. (1962). On the social psychology of the psychological experiment with particular reference to demand characteristics and their implications. *American Psychologist*, 17, 776–783.
- Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *Journal of Experimental Education*, 58, 311–320.
- Partridge, E. (1966). *Origins* (4th ed.). New York, NY: Macmillan.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY: Cambridge University Press.
- Pearl, J. (2018). *The book of why: The new science of cause and effect*. New York, NY: Basic Books.
- Peters, D. P., & Ceci, S. J. (1982). Peer review practices of psychological journals: The fate of published articles, submitted again. *The Behavioral and Brain Sciences*, 5, 187–195.
- Piaget, J. (1950). *Introduction à l'épistémologie génétique [Introduction to genetic epistemology]*. Paris, France: Presses Universitaires de France.
- Pirsig, R. (1974). *Zen and the art of motorcycle maintenance*. New York, NY: William Morrow.
- Plooij, F. X. (2020). The phylogeny, ontogeny, causation and function of regression periods explained by reorganizations of the hierarchy of perceptual control systems. In W. Mansell (Ed.), *The interdisciplinary handbook of Perceptual Control Theory: Living control systems IV* (pp. 199–225). London, UK: Elsevier.
- Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités [Investigations of the probability of judgments in criminal and civil matters, preceded by general rules of the calculus of probabilities]*. Paris, France: Bachelier.
- Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy* (Corrected ed.). Chicago, IL: University of Chicago Press.
- Pollard, P. (1993). How significant is “significance”? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 449–460). Hillsdale, NJ: Erlbaum.
- Porter, T. M. (1992a). Objectivity as standardization: The rhetoric of impersonality in measurement, statistics, and cost-benefit analysis. *Annals of Scholarship*, 9, 19–59.
- Porter, T. M. (1992b). Quantification and the accounting ideal in science. *Social Studies of Science*, 22, 633–652.
- Powers, W. T. (1973). *Behavior: The control of perception*. Chicago, IL: Aldine.
- Powers, W. T. (1978). Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology. *Psychological Review*, 85, 417–435.
- Redner, H. (1987). *The ends of science: An essay in scientific authority*. Boulder, CO: Westview Press.
- Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 445–462). Thousand Oaks, CA: Sage.
- Rosenbaum, R., & Dyckman, J. (1995). Integrating self and system: An empty intersection. *Family Process*, 34, 1–23.
- Rosenthal, R. (1963). On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, 51, 268–283.
- Sampson, E. E. (1981). Cognitive psychology as ideology. *American Psychologist*, 36, 730–743.
- Sampson, E. E. (1985). The decentralization of identity: Toward a revised concept of personal and social order. *American Psychologist*, 40, 1203–1211.

- Sampson, E. E. (1988). The debate on individualism: Indigenous psychologies of the individual and their role in personal and societal functioning. *American Psychologist*, 43, 15–22.
- Samuelson, P. A. (1964). *Economics* (6th ed.). New York, NY: McGraw-Hill.
- Savić, D., & Jelić, S. (2005). A mathematical model of the hypothalamo-pituitary-adrenocortical system and its stability analysis. *Chaos, Solitons & Fractals*, 26, 427–436.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of statistics? *Psychological Bulletin*, 105, 309–316.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.
- Shafer, G. (1990a). The unity and diversity of probability. *Statistical Science*, 5, 435–462.
- Shafer, G. (1990b). The unity of probability. In G. M. von Furstenburg (Ed.), *Acting under uncertainty: Multidisciplinary conceptions* (pp. 95–126). Boston, MA: Kluwer Academic.
- Shafer, G. (1993). Can the various meanings of probability be reconciled? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 165–196). Hillsdale, NJ: Erlbaum.
- Shafer, G. (1996). *The art of causal conjecture*. Cambridge, MA: MIT Press.
- Shapiro, D. (1965). *Neurotic styles*. New York, NY: Basic Books.
- Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, 61, 293–316.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.
- Skousen, M. (1997). The perseverance of Paul Samuelson's *Economics*. *Journal of Economic Perspectives*, 11, 137–152.
- Smith, M. B. (1994). Selfhood at risk: Postmodern perils and the perils of postmodernism. *American Psychologist*, 49, 405–411.
- Spencer Brown, G. (1972). *Laws of form*. New York, NY: Julian Press.
- Spiegelberg, H. (1982). *The phenomenological movement: A historical introduction* (3rd ed.). The Hague, The Netherlands: Martinus Nijhoff.
- Star, S. L. (1989). *Regions of the mind: Brain research and the quest for scientific certainty*. Stanford, CA: Stanford University Press.
- Stone, L. (1977). *The family, sex and marriage in England 1500–1800*. New York, NY: Harper & Row.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: SAGE.
- Student. (1931). The Lanarkshire milk experiment. *Biometrika*, 23, 398–406.
- Thompson, M. (1979). *Rubbish theory: The creation and destruction of value*. Oxford, UK: Oxford University Press.
- Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358–376.
- Tripp, C. A. (1975). *The homosexual matrix*. New York, NY: McGraw Hill.
- Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
- von Eckartsberg, R. (Ed.). (1986). *Life-world experience: Existential-phenomenological research approaches in psychology*. Washington, DC: University Press of America.
- Wann, T. W. (Ed.). (1964). *Behaviorism and phenomenology*. Chicago, IL: University of Chicago Press.
- Wartofsky, M. W. (1968). *Conceptual foundations of scientific thought*. New York, NY: Macmillan.
- Wartofsky, M. W. (1991). How to be a good realist. In G. Munévar (Ed.), *Beyond reason: Essays on the philosophy of Paul Feyerabend* (pp. 25–40). Dordrecht, The Netherlands: Kluwer Academic.
- Weisberg, H. I. (2014). *Willful ignorance: The mismeasure of uncertainty*. New York, NY: Wiley.
- Williams, R. J. (1956). *Biochemical individuality: The basis for the genetotrophic concept*. New Canaan, CT: Keats Publishing.

- Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York, NY: McGraw-Hill.
- Yin, H. H. (2014). Action, time and the basal ganglia. *Philosophical Transactions of the Royal Society B*, 369. <https://doi.org/10.1098/rstb.2012.0473>
- Young, I. M. (1990). The ideal of community and the politics of difference. In L. J. Nicholson (Ed.), *Feminism/postmodernism* (pp. 300–323). New York, NY: Routledge.

Index

Numbers

0–9

.05 level origins, 146–153w

A

Accad, M., 26, 401

Moving Mountains, 26, 431

Ancillon, J.P.F., 109, 394

Application to psychological research,
problems in

assumption violation, 320

epistemic *vs.* behavioral

orientation, 310–311

identification with the null, 316–317

individual *vs.* aggregate, 312–314

literalness of acceptance, 311–312

paradox of precision, 314–316

Arbuthnot, J., 62, 143–145

B

Bacon, F., 50, 52, 53, 66, 110, 153

Bayes, definition of probability, 89

Bayesian inference, 10–11, 155, 207,
260–266, 304–305, 321–328,
340–345, 375–378, 381–382,
385–386, 398–399

Benjamin, J., 68, 413, 420, 425, 429

Berman, M., 34, 412

Bernard, C., 138, 146

Bernoulli, D., 90, 146, 147, 192, 287, 362

Bernoulli, J., 13, 23, 94–98, 100, 102–109,
112–113, 117–123, 137, 143, 145,
160, 165, 170, 243, 315, 385, 393

Bertrand, J., 137, 140, 162, 214, 248

Biomedical research without a biomedical
model, 431–433

Biometrics, 139–143

Bordo, S., 34, 58–59, 64, 69

C

Causality, concept of
changes in, 51–55

Change

deep obstacles

Dinnerstein's theory applied to
epistemology, 416–420

intersubjectivity and self-
responsibility, 420–425

paradoxes of control, 420–425

research as a cultural obsessive-
compulsive disorder, 412

for quantitative research, 405

role of effect sizes, 406–408

for qualitative methods

reducing the data *vs.* reducing the
question, 408–409

Giorgi made phenomenology the study
of interviews, 410–411

surface obstacles 400–404

government funding, 400–402

need for separation of state and
science, 402

peer review, 403

questions have to change with
methods, 404

Chauvenet, W., 147, 196

Childhood, disowning of, 68–72

- Cohen, J., 5, 8, 16, 18, 315, 406, 407
 Cohen, L.J., 110, 243, 336, 343, 344, 365,
 370, 371, 398, 399
 Combination of evidence, 103
 Combinatorics of language and thought, 49–51
 Confidence intervals
 backward look, 226
 Corbett, J., 19, 20
 Cournot, A.A., 112, 113, 137, 175–176, 239
- D**
 d'Alembert, J. le R., 68, 100, 110, 179, 180
 Danziger, K., 25, 287, 292, 294–297
 Daston, L., 46–47, 65, 80, 81, 89, 91, 112
 de Finetti, B., 127, 237, 250, 251, 253–257,
 271, 272, 274–276, 305, 363
 de Moivre, A., 68, 95, 108, 132
 Descartes, 50, 52, 53, 56, 64, 68–70, 189
 Dinnerstein, D., 413, 416–418, 429
- E**
 Easlea, B., 63, 66, 67
 Economy, market
 growth of, 43–45
 Edgeworth, F.Y., 14, 140, 150, 152, 153, 283
 Edwards, A.W.F., 208, 341, 363–366
 ESP research, 248, 308
- F**
 Farley, L., 26
 Feminine
 rejection of the, 64–68
 Feyerabend, P., 16, 306, 370, 384, 409, 411, 429
 Fiducial probability, 203–209
 recent efforts to rehabilitate, 207
 Fine, T., 181, 182
 Fisher, I., 283–284
 Fisher, R., 11, 22, 150, 172, 180, 187–209,
 260, 267, 281, 293, 295, 297, 302,
 306, 319, 362–363, 397, 398
 Formalization
 limits, 379–382
 purposes, 383
 Foucault, M., 40–41, 48, 49, 55, 70, 123, 397,
 416, 421
 Franklin, J., 53, 66, 80, 83–84, 88
- G**
 Galton, F., 22, 139–143, 151, 293
 Gavarret, J., 145–146, 152, 385
 Gerson, E., 25
- Gigerenzer, G., 25, 305, 339, 341, 343–345,
 353, 400, 405, 406
 Goldberg, S., 2
 Gosset, W.S., 153–156, 180, 188, 196, 202,
 211, 296, 404
- H**
 Hacking, I., 13, 49, 53, 80, 83, 87, 90, 94, 98,
 106, 107, 119, 128, 198, 217, 219,
 247, 359, 363
 Hays, W.L., 12, 220, 228, 230, 260, 261, 263,
 264, 273, 304, 305, 315, 326, 406, 407
 Hierarchical models, 326–327
 Hogben, L., 5, 7, 122, 142, 152, 180, 199,
 220, 224
 Hornstein, G., 25, 281, 287–289, 291
 Huygens, C., 84, 88–89, 143
 Hypothetical infinite population
 circularity of, 197, 198
- I**
 Indifference, principle of, 124–129, 161–164
 Inference
 atmospheric effects, 347–350
 models
 Bayes' Theorem, 340–345
 causal inference and ANOVA, 347
 clinical inference and multiple
 regression, 345–347
 statistical and nonstatistical, 335–338
 Ioannides, J., 18
- J**
 Jaynes, E.T., 223–224, 248–250, 268, 313
 Jaynes, J., 35, 39
 Jeffreys, H., 15, 53, 148, 217, 219, 237, 244, 250,
 261–263, 267, 268, 275, 305, 382
- K**
 Kahneman, D., 338, 341–343, 346, 348, 352
 Kaye, J., 41–45, 282
 Keller, E.F., 56, 66–68
 Kendall, M.G., 4, 12, 16, 33, 68, 71, 154, 180,
 188, 197, 223, 231, 275, 393
 Keynes, J.M., 98, 100, 110, 111, 124, 126,
 128, 134, 152, 164, 214, 237–245,
 252, 270, 272, 284–286, 382, 399
 disparaged mathematical
 economics, 284–286
 not all probabilities measurable, 241–244
 Knight, C., 3, 42, 43

- Koch, S., 39, 308, 427
Koestler, A., 3, 57, 378, 417
K, posterior odds ratio, 262–263
Kyburg, H., 191, 207, 208, 226, 251, 270, 273, 323, 360
- L**
Lambert, J.-H., 13, 104–108, 192, 362
Langer, S., 6, 384–385, 420
Laplace, M. de, 101, 109, 111, 118, 122, 124–125, 131–134, 136, 141, 142, 396
Large numbers, law of, 96
Latitudo
neglected concept from medieval price theory, 282
Likelihood theory of statistical inference, 362
Likert, R., 289
Lindley’s paradox, 265
Loevinger, J., 290
Lonergan, B., 336–338
- M**
Maximum likelihood, 192–194
Measurement
errors in astronomy, 132–133
in psychology, 281–294
Medicine
statistics in, 328–329
Mises, R.v., 155–156, 163, 168–173, 177, 194, 214, 237, 274, 276
Multiple imputation of missing data, 327–328
- N**
Neuropsychology, Bayesian, 386–387
Neyman and Pearson, 172, 187, 209–231, 259, 260, 298–301, 303, 304, 310, 312, 317, 340, 341, 345, 362, 365, 369, 398, 399, 405
Neyman, J.
equivocation on probability, 214–217
Normal curve
application to populations, 134–139
development of, 131–134
Nosek, B., 18
Nozick, R., 28, 426, 427
- O**
Oakes, M., 5, 8, 11, 14, 221, 266, 270, 275, 322, 352, 353
Objectivity, scientific, 46–47
vs. intersubjectivity, 420–430
- P**
Pascal, B., 58, 60, 68, 87, 90
Pearl, J., 371, 373, 374, 376–379
Pearson, K., 22, 61, 67, 68, 86, 92–94, 112, 131, 137, 139, 141, 144, 154, 188, 189, 191, 195, 196, 209, 211, 281, 293
Peirce, C.S., 164, 166–168, 177, 358
Perceptual Control Theory, 434–437
Petersburg game, 90, 253
Piaget, J., 22, 24, 127, 163, 164, 336, 354, 356, 357, 381, 411, 422
Poisson, S.-D., 76, 112, 125, 395
Polanyi, K., 3, 382, 401
Popper, K., 164, 169, 173, 176–178, 274, 358–362
Porter, T., 46, 131, 137, 140, 149, 152
Port-Royal *Logic*, 48–49, 83, 88
Powers, W., 380, 413, 434
Printing press, 47
Probability, concept of
definition, 9
everyday, 9
future of, 393–395
incoherence, 9–10
metaphysical status, 99–100
ontogenesis, 354–357
origins, 33, 79–91
two scales, 101–107
Propensity theory of probability, 358–362
- Q**
Quetelet, A., 125, 134–139, 169
- R**
Ramsey, F., 237, 251–256
Randomization tests, 14, 198–201
Randomness, concept of, 178
judgments of, 350–351
of samples, 317–320, 351–354
Rationalism
grip in nineteenth century, 159–160
Reichenbach, H., 172–176, 184, 274, 276, 336, 356
Replication problem, 18–20, 351–354
Rule of Succession, 124–129, 161–164, 178, 244, 295, 304
- S**
Savage, L.J., 256–258, 274, 363
Scott, J., 92, 135

- Self-consciousness, 39–41
growth of
 Greece, 35–37
 Medieval Europe, 37–39, 58–60
- Shafer, G., 13, 23, 88, 95, 96, 101, 104, 105,
 107–109, 119, 243, 265, 267,
 365–371, 375–380, 393–395,
 398, 399
- Sign, concept of
 changes in, 47–49
- Skeuomorphosis, 427–428
 defined, 41–42
- Small samples
 liability turned to advantage, 194–202
- Social class and ideology, 27
- Spencer Brown, G., 21, 26, 179–182, 314–315,
 339, 351, 384, 418, 419, 422, 423
- Statistics
 as science of the state, 91–92
- Stigler, G., 98, 118, 122, 135, 140, 146,
 149–151, 286
- Subjectivity
 of Bayesian methods, 272–276
 formalizing *vs.* eliminating, 273, 323–326
- Sylla, E., 49, 87, 96, 98, 101,
 103, 107
- Symbol Formation*, 51, 373, 386
- T**
- Task Force on Statistical Inference,
 APA, 5
- Taylorism, 291–298
- Testimony, probability of, 110
- Theory and practice disparity
 Fisher and Jeffreys, 277
- Thurstone, L., 288
- Tversky, A., 338, 341–343, 346, 348, 352,
 368, 380, 405
- V**
- Velikovsky, I., 35, 307
- W**
- Wald, A., 254, 258–260, 264
- Wolfram, S., 18, 338