Exploring the History of Statistical Inference in Economics: Introduction

Jeff Biddle and Marcel Boumans

We had two motivations that made us decide that an exploration of the history of statistical inference could be productive. One was related to John Maynard Keynes's distinction between two different functions that he observed in statistical research, and the other is that existing histories of empirical analysis seem to miss an important part of existing research practices in economics.

Keynes, in his *Treatise on Probability*, made a distinction between the descriptive function of the theory of statistics, which involved devising ways to represent and summarize large amounts of data, and the inductive function, which "seeks to extend its descriptions of certain characteristics of observed events to the corresponding characteristics of other events that have not been observed" (Keynes [1921] 1973: 359). This second part of statistics he called the theory of statistical inference. When looking at any given example of statistical research in economics, one is likely to see both what Keynes called description and what he called inference; indeed, it is not always obvious where one ends and the other begins. The researcher will have made decisions about how best to summarize statistical data, and

We would like to thank all participants of the conference at which these papers were discussed for the first time. Besides the contributors to this volume, they were Kevin Hoover, Jennifer Jhun, and Steven Medema. We are very grateful to our external reviewers for their valuable and constructive reports: Maria Bach, H. Spencer Banzhaf, Beatrice Cherrier, Erwin Dekker, Federico D'Onofrio, Verena Halsmayer, Catherine Herfeld, Steven Medema, Theodore Porter, and Gerardo Serra.

History of Political Economy 53 (annual suppl.) DOI 10.1215/00182702-9414747

also decisions about what generalizations can be made, with what level of confidence, based on the information in the sample. These two sets of decisions are often related: the choice of how to summarize the sample information is influenced by beliefs about which summary measures provide the firmest bases for the types of generalizations the researcher hopes to make. But Keynes's distinction is a real and important one, nonetheless. Crossing the line from description of the information in a sample to inference about things beyond the sample, no matter how inconspicuous the act of crossing may be, necessarily involves making assumptions, implicitly or explicitly, about the relationship between the data in the sample and the phenomena outside the sample about which generalizations are being made. If one accepts that Keynes's distinction is a meaningful one, it follows that bound up with the history of statistical analysis in economics, there is a distinguishable history of statistical inference, a history of the ways in which economists have gone about generalizing from statistical data.

The second motivation can also be clarified by a distinction. Edward Leamer (1978: vi) observed a wide gap between the formal textbook approach, "taught on the top floor (the third)," and its practiced variant, "done in the basement of the building": "I was perplexed by the fact that the same language was used in both places. Even more amazing was the transmogrification of particular individuals who wantonly sinned in the basement and metamorphosed into the highest of high priests as they ascended to the third floor." The people in the basement are sinners because their research does not meet the high standards of the discipline, often presented more generally as the standards of science. It seemed to us that most histories are written about the good works of these high priests, with little attention to the works of Leamer's "sinners."

There may be a reason why we do not have so many histories of sinful research. It is smudgy and messy and therefore does not lend itself to "seamless accounts to make it comprehensible," for which one has to "paper over the knots and holes in scientific life" (Morgan 2012: xv). More generally, the study of research practices faces distinct and substantial challenges. In a reflection on practice-oriented studies, which often are case studies, Andrea Woody discusses two of these challenges. The first, obvious challenge in case study research is how to generalize from a particular case, "how to avoid getting stuck at a level of particularity that evades any reasonable effort to generalize" (Woody 2014: 124). The more vexing challenge, however, has to do with normativity, that is, the issue of the abovementioned sinning. In our assessments of a practice, we build on

strong intuitions and a priori reasoning: "In effect, the normativity of the analysis is built in from the get-go" (125). Although Woody focuses only on philosophical assessments, we believe that this built-in normativity also restricts the kinds of practices that are studied by historians. A priori concepts of "good science" influence the selection of historical cases of economic research that historians find worthwhile to investigate. In this sense, history of economic science reflects contemporaneous philosophy of science.[1] Tjalling Koopmans's (1947) negative review of the work done at the National Bureau of Economic Research (NBER), which he criticized as "measurement without theory," did not only set the scientific standards of empirical research in economics for decades; it also blinded historians to many practices that did not meet Koopmans's standards for proper research (see also Stapleford, this volume). The result is that there are relatively few historical studies of the work of Irma Adelman, Wassily Leontief, or Simon Kuznets, even though the latter two are Nobel laureates.

A number of interesting narrative threads, involving varied and changing inferential methods and strategies, can be discerned in the history of empirical economic research. One that has gotten a fair amount of attention (although mostly indirectly, in the context of histories of econometrics) involves the process through which the use of inferential procedures derived from probability theory, and designed to measure the possible impact of sampling error on the reliability of statistical estimates, came to be ubiquitous in empirical research in economics.

To be more specific, by the last decades of the twentieth century, a broad consensus had developed in the economics profession that statistical inference required the use of these procedures. A rather recent econometric textbook, *Econometric Analysis* (Greene 1999) even labeled this consensus "classical theory of inference," thus suggesting that this consensus is old and lasting, which is an important feature of standards. According to this standard for statistical inference, a plan for data analysis typically begins with a set of assumptions about the joint distribution of random variables of interest in a population of interest. This joint distribution is characterized by fixed parameters that embody important but

1. This normativity was explicitly vindicated by Imre Lakatos, aptly summarized by his dictum that "history of science without philosophy is blind." While this view on history today is only embraced by a very few historians, we nevertheless believe that scientific norms and standards still play an important role in the selection of what is interesting or relevant to study. For a detailed discussion of the relationship of the image of science and history of econometrics, see Boumans and Dupont-Kieffer 2011.

unknown facts about some phenomenon of interest. The assumptions are usually patterned on one of a set of canonical "statistical models" with well-understood properties, such as the linear regression model.

It is also assumed that the set of observations of the relevant variables is a *random sample* of observations taken from the population of interest. Statistical inference is, then, a matter of applying formulas to this sample information. A formula produces estimates of the parameters, and other formulas produce measures of the reliability of those estimates—"standard errors" and so forth. These formulas have been derived from probability theory, and there is a set of procedures—also derived from probability theory—for using the estimates produced by the formulas to test hypotheses about the parameters.

The work of Trygve Haavelmo and the Cowles Commission econometricians, in particular Koopmans, in the 1940s and early 1950s was crucial to the eventual widespread acceptance by economists of this approach to statistical inference. The Cowles group's justification of the application of probability theory to economic data is found in Haavelmo's "Probability Approach in Econometrics." Koopmans (1947) based his "measurement without theory" attack on the nonprobabilistic approach to the analysis of business cycles of Arthur F. Burns and Wesley C. Mitchell (1946) on his interpretation of the major message of Haavelmo's 1944 "probability approach," namely, that for *scientific* statistical inference, probability theory is essential. By the early 1970s, a standardized set of inferential methods justified by probability theory—methods of producing estimates, of assessing the reliability of those estimates, and of testing hypotheses—was being taught to the majority of economics graduate students. The phrase *statistical inference* had come to mean, in the minds of economists, the application of a set of techniques derived from probability theory (Biddle 2017), and for many economists the term *econometric* was reserved for empirical research that employed these methods and the methods of estimation that they accompanied (Biddle 2021: 280).

There is a robust historical literature about the activities of Haavelmo, Koopmans, and other early Cowles Commission econometricians and the subsequent development of the approach to statistical analysis that by the late twentieth century had inherited the title "econometrics" (Boumans, Dupont-Kieffer, and Qin 2011; De Marchi and Gilbert 1989; Epstein 1987; Morgan 1990; Qin 1993, 2013). This literature necessarily deals with the development of inferential methods based on probability theory. However, one of our motivations for organizing this conference is that the "history

of statistical inference" that emerges from this literature leaves out much that is important and interesting.

For example, despite the fact that discussions of some of the principles and procedures of the "classical theory of inference" could be found in standard textbooks on economic statistics in the 1920s, prior to World War II very few empirical economists in the United States made any use of these tools of statistical inference to draw conclusions from statistical data. In the immediate postwar decades, as the use of the inferential tools derived from probability theory was coming to be regarded as the scientific standard for empirical research in economics, nonprobabilistic approaches to inference persisted in the common practice of empirical economics, and new nonprobabilistic approaches to inference were being developed. These modes of inference were in a sense put on the defensive, however, by the increasing popularity of the "econometric" approach. They became less visible, and we believe less studied, by historians of economics partly because many of the economists who practiced and taught them were not working on Leamer's "top floor" but in the "basements" of the academic departments of economics, and often outside academics altogether in government or the private sector,[2] and also because the dominance of the probabilistic approach was confused with its dispersion. It was assumed that coverage of the probabilistic approach would be sufficient to understand the history of postwar empirical research.

Thus, as we began to organize the conference, we were hoping for contributions that would bring these practices more to the foreground and show how widespread they actually were in the latter decades of the twentieth century, as well as contributions that would throw light on the many nonprobabilistic approaches to inference found in the work of economists writing before 1940. We were not disappointed.

With the support of Duke University Press and the editors of this journal, we invited a number of scholars with an interest in the history of empirical research in economics to contribute papers on episodes in or aspects of that history that would highlight the various ways in which Keynes's "statistical inference"—generalizing from evidence provided by a set of statistical data to statements about phenomena not described in those data—manifested itself. We particularly invited them to explore those practices that did not meet Koopmans's standards.

2. Histories of nonprobabilistic approaches can be found in *HOPE* conference volumes covering more general themes and a longer time period, such as Klein and Morgan 2001 and Maas and Morgan 2012.

The conference was scheduled for April 2020. By March, however, it became apparent that because of the COVID-19 pandemic, a traditional style of conference should not be held, so a "virtual" conference was planned instead. We anticipated with some disappointment the loss of certain attractive features of traditional conferences: the spontaneous and informal discussions between sessions and over meals of topics and questions raised during the formal sessions, and the chance to renew old acquaintances and talk about shared interests. And these things were indeed missed. But the conference sessions, though mediated by internet technology and involving participants in locations (and time zones) from Berlin to Berkeley, were orderly and stimulating. Further discussions between participants took place by email between the closing of one day's sessions and the opening of the next and continued in the days after the conference as participants turned to the task of revising their papers. There was wide agreement that the conference had worked out well.

The conference papers and discussions provided strong indications of the existence of a promising area for research. There was a general consensus that attention to inference as an analytically separable aspect of research involving statistical data could be a fruitful perspective from which to understand the history of empirical economics. Throughout the conference, participants introduced, and elaborated through discussion, several potentially useful ways to understand "statistical inference" and conceptual frameworks for thinking about it. And the papers themselves provided an indication of the fascinating variety of topics and themes that could ultimately be comprehended as part of the history of inferential methods and practices in economics.

The resulting chapters can be grouped together by three general themes: Inferences in the Field (Burnett, Biddle, and Samuel), Inference in Time (Morgan, Lenel, Stapleford), and Inference without a Cause (Boumans, Velkar, Akhabbar, and Maas). The concept of field has several connotations that are relevant for the cases studied in the first three chapters. It refers to research not done in Leamer's building at all but outside academic institutions and universities. Field, of course, can also be taken literally, and the chapters of this first section discuss research in agricultural economics and economic development. The concept of time plays a central role in the cases of the next three chapters but in two different ways. Mary S. Morgan and Laetitia Lenel discuss research on economic developments in time, such as trends and cycles, while Thomas A. Stapleford discusses the development of statistical research in time. The final four

chapters discuss inferences made with a cause, but where the achievement of this cause or the cause itself was questioned.

Thinking about Statistical Inference: Themes Emerging from the Conference

The conference proposal shared with the contributors was centered on Keynes's 1921 definition of inference and his opposition between statistical description (within a sample) and statistical inference (making statements about things not observed in the sample on the basis of the information in the sample). As drafts of papers circulated and discussions began, however, it became clear that an idea of inference based on Keynes's presentation was, for several reasons, unduly narrow.

First, Keynes's concept of statistics was a twentieth-century concept. As G. U. Yule (1911) explained, the meaning of *statistics* had evolved over the eighteenth and nineteenth centuries, originally referring to expositions, largely verbal, of noteworthy characteristics of a state. Only gradually did the word *statistics* come to be associated with numerical information, lose its close association with information about a state, and acquire the meaning that Yule and Keynes attributed to it. So, as a historical matter, the evidence with which statistical inference has dealt is not limited to the numerical information that Keynes had in mind.[3] Accordingly, in this volume, we have Morgan's account of Thomas Robert Malthus's use of narrative in making inferences from evidence that was *statistical* in the eighteenth-century sense of the word, while Marcel Boumans describes Francis Galton's project of applying concepts from nineteenth-century statistics to draw inferences from photographs, based on the assumption that a collection of photographs will have some of the same properties as the collections of measurements that make up conventional statistical samples.

Second, Keynes's distinction between description and inference leaves out what is a necessary element of any statistical project in economics, the collection of the statistical data. And just as the decisions about how to describe data already collected are often influenced by a consideration of what sorts of inferences one hopes to make using the data, so are decisions about the data collection process—how concepts will be defined, how they will be measured, the composition of the sample, and so on. In

3. According to Yule (1911: 3), at the founding of the Royal Statistical Society in 1834, *statistics* was still being used to describe both numerical and nonnumerical information.

Harro Maas's chapter in this volume, for example, one sees how debates over the credibility of inferences presented in contingent valuation studies often focused on details of how the data were gathered, while Jeff Biddle's chapter describes how the US Department of Agriculture's economists reacted to past errors in forecasting crop and livestock production by redesigning data collection instruments and sampling procedures.

Third, Keynes's discussion of inference often referred to the "sample-universe" framework for thinking about inference: if one had a sample of statistical data, how could one best use it to make generalizations about some universe or population of interest beyond the sample? This way of thinking required the investigator to explore whether the members of the sample in hand had actually belonged to the universe of interest, how "representative" of the universe the sample was, in what specific sense a sample might not be representative of the universe, and so on. The sample-universe conceptualization was part and parcel of the project of developing techniques of statistical inference based on probability theory and is almost taken for granted in modern discussions of statistical inference. But it was just becoming familiar to empirical economists at the time Keynes wrote. Yule (1911) made use of it in his *Introduction to the Theory of Statistics*, and it would be discussed in the leading books on statistical methods for economists published in the United States in the 1920s (Biddle 2017). However, one finds in economists' writings of that and the next several decades ample evidence of a belief that the sample-universe framework was not a good one for thinking about how to draw credible inferences from many of the types of data with which empirical economists were working. The chapters by Biddle and Laetitia Lenel in this volume discuss the strong opinion of economists in the 1920s and 1930s that it was not useful to regard time series as samples from a universe, especially when forecasting was the goal of inference. Amanar Akhabbar (this volume) shows that Leontief made little use of sample-universe thinking in developing inferential procedures for his early input-output analyses, and he was unmoved by the opinion of econometricians, whose "indirect" methods of inference he criticized, that he would do better to reconceive his project of estimating input-output coefficients in a sample-universe framework. And much of the statistical material used in national income accounting and related efforts to describe national economies still does not lend itself to a sample-universe way of thinking, as is clear from Boris Samuel's chapter in this volume.

As noted above, conference discussions led to the elaboration and refinement of several potentially useful heuristics and conceptual frameworks for historians examining the role and nature of statistical inference in empirical economics. A first is a simple observation by Morgan that inference is a verb as well as a noun and that one could be interested in understanding a process or in the outcome of that process. In her chapter in this volume, Morgan observes that the terminology of inference nowadays seems to just refer to the outcome, not the process, of drawing the inference from the evidence. Halfway through the twentieth century, the various informal and tacit practices of inferences came to be replaced by more formal statistical inference based on explicit rules and procedures with clear criteria to ensure the reliability of the process. It seems that when such processes became rule bound, the process of making the inference and the statement of the inferential outcome somehow became conflated, with the consequence that the process of inference became less visible.

It is valuable to develop a clear understanding of the question or goal that motivated an inferential process, who was asking it, and why. Failure to do so risks missing the forest among the trees that are the often complicated statistical and inferential procedures that economists employ. Aashish Velkar's chapter in this volume produces interesting insight into the nature of statistical inference by looking at the use of the same basic descriptive statistical technique—the construction of a price index number—to answer two different questions: Stanley Jevons's attempt to measure changes in the value of the monetary standard versus the British Board of Trade's efforts to measure changes in workers' cost of living. The question behind the inferential activities described in both Biddle's and Lenel's chapters is easily grasped—the economists were making forecasts of future economic activity, and they adjusted their inferential procedures over time as their forecasting errors were revealed. But in two of the chapters, the relationship between the question motivating the research and the process and outcome of inference are less straightforward. In Boumans's chapter, we see that Galton's method of handling photographs did not allow him to make the sorts of inferences he had hoped to make about the facial characteristics of racial or criminal types, but it did surprise him by pointing to an inference about a different question: the relationship between facial "beauty" and the individual irregularities of appearance among the photographed subjects. In Samuel's chapter, we see that an accurate description of past macroeconomic conditions or forecasts of

future ones for the purposes of recommending policy was only one of many goals sought by International Monetary Fund economists in their work with data, with the result that they adopted inferential practices that they understood to be less effective than others that they could have used.

The idea of inferential gaps, an expansion of a concept of "inferential distance" employed by Kevin D. Hoover and Michael Dowell (2001), seemed to the conferees a promising one. In one sense, inference is necessitated by the existence of *a* gap, that between the statistical data one has in hand, known with complete certainty, and the phenomenon about which one wants to generalize, which is unknown. This gap can also be felt to be very large (the "distance" can be very long), in particular when the phenomenon is not part of our daily experiences and hence is conceived of as "strange," as Burns and Mitchell (1946: 17) described the result of their inferences that led to a conceptualization of the business cycle:

> Thus the concept of business cycles ties together in our minds, and gives meaning to, a host of experiences undergone by millions of men, few of whom think of themselves as influenced by cyclical pressures and opportunities. The concept, as we develop it, is itself a symbol compounded of less comprehensive symbols representing the cyclical behavior characteristic of many unlike activities. In turn, these symbols are derived by extensive technical operations from symbolic records kept for practical ends, or combination of such records. We are, in truth, transmuting actual experience in the workaday world into something new and strange.

Inference is, in a sense, the attempt to bridge this gap; it is what happens "between evidence and expression," as Stapleford put it at the conference. But a sense developed at the conference that it was perhaps better to be on the lookout for many types of inferential gaps that might arise in an empirical research project and the strategies employed to bridge them in the process of creating plausible inferences.

An analogy with a bridge, however, can be misleading with respect to the problems that the empirical researchers were facing. A bridge goes from *known* to *known*. For Haavelmo (1944: iii), it was "a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier." His "Probability Approach in Econometrics" aimed at building such a bridge. But often the other side of an inferential gap could not be seen.

Gaps can originate for various reasons. Gaps arise from data quality problems, when the measures available to the researcher of the things he or she wants measured are inaccurate or imprecise. There is concept mismatch: the thing that was measured (well or badly) is not exactly the thing that the researcher wanted to make inferences about. There is sample-universe mismatch: the sample is representative of one environment, but the researcher wanted to draw an inference about a different environment. And there is sampling bias and sampling error in samples that come from the environment the researcher cared about. In Velkar's contribution to this volume, one sees still another inferential gap: that between the inferences about the cost of living to which the available data and the accredited statistical procedures lead the experts and the inferences of ordinary citizens based on their experiences—a gap that involves the strangeness referred to in the Burns and Mitchell quotation above. Having identified the inferential gaps that faced a researcher or researchers, it is then worth asking how much attention was given to each of these problems and what steps were taken to solve them.

It was suggested that when considering the history of statistical inference more broadly, one might be able to identify certain inferential strategies common to a number of historical episodes. A familiar example is the strategy of trying to quantify the extent to which a sample statistic, say a regression coefficient, might differ from the unknown population value it is meant to estimate. The tactics associated with this strategy are based on probability theory and taught as a matter of course to aspiring economists. But other common strategies have long been employed by economists. Maas and Morgan discuss the strategy of *triangulation*. The strategy is based on the idea that we can have more trust in an inference when information derived from different methods and sources are found to be congruent and consistent with the same conclusion. Inferences based on descriptive statistics of a certain sample of data may be bolstered by evidence from other types of samples and statistical materials, from interviews, from surveys, and so forth. The agricultural economists in Biddle's paper "triangulated" by comparing the forecasts of production they derived from the reports of volunteers in the agricultural areas with the later reports of railroads about volumes of freight carried. Paul Burnett's chapter in this volume describes how Zvi Griliches used interviews with seed company executives to make sense of the shapes of empirical curves derived from data on hybrid corn adoption. The triangulation tactic of interviewing experts as an aid to drawing good inferences also appears in

Lenel's account of the Harvard Economic Service's increasing reliance on the opinions of bankers and businessmen along with its statistical model in developing forecasts.

Another inferential strategy involves the use of theory and assumed theoretical relationships to go from observed and measured phenomena to unobserved phenomena of interest. Using data on trends in wages along with the marginal productivity theory of distribution to infer trends in productivity would be one example, and Samuel (this volume) describes how IMF economists use the quantity theory of money and the monetary approach to the balance of payments to derive estimates of unmeasured macroeconomic quantities from the data available to them.

Morgan's contribution to the volume makes the intriguing proposal that the construction of narratives might be a process through which researchers arrive at inferences, which would make narrative an inferential strategy worth looking for in the history of empirical economics. Burnett's chapter on Theodore Schultz's (1964) discussion of statistical evidence in *Transforming Traditional Agriculture* provides a fascinating analysis of what may be another general inferential strategy. He describes how Schultz assembled and interpreted a small assortment of statistical studies, making each, with the help of basic neoclassical theory, into a "statistical parable" that supported his arguments in a controversy with other development economists in the late 1950s and early 1960s.

Toward a Broader View of the History of Statistical Inference in Economics

What follows is a selective sketch of the history of statistical inference in economics in the twentieth century, meant to place the better-known narrative thread centered on the development and growing prominence of inferential tools derived from probability theory into the wider context provided by the variety of approaches to statistical inference being used throughout the period. In doing so we highlight the contributions of the chapters in this volume to fleshing out that broader history and suggest some further opportunities for research.

Although this sketch focuses on the twentieth century, statistical inference as an activity of economists is older than that, as Morgan, Boumans, and Velkar remind us. At the turn of the nineteenth century, Malthus was engaging in statistical inference when "statistics" meant something altogether different than it does now. At the end of the 1800s, Galton was

experimenting with ways to draw "statistical" inferences from collections of photographs. By the early twentieth century, the price index number was a well-enough-established statistical device that the government of the United Kingdom was maintaining a cost-of-living index, which became the basis of conflicting inferences about trends in British living standards and appropriate economic policies. Probability theory, however, had no significant role in the economic approaches of the nineteenth and early twentieth century. It was believed that probabilistic laws governed only errors and deviations, and the methods and procedures of inferences that were used were based on this belief. In other words, the methods were designed to deal with uncertainty in terms of ignorance; in economics the deterministic worldview was still dominant.

During the first two decades of the twentieth century, there were a few economists (e.g., Arthur Bowley [1920], F. Y. Edgeworth [1913], Yule [1911]) who explained how probability theory could be used to generalize from statistical data. However, the measures associated with these inferential procedures (e.g., the *probable errors* of means and correlation coefficients) are rarely seen in the statistical work of the time. Keynes's 1921 *Treatise on Probability* included an explicit and detailed rejection of probability theory as a basis for statistical inference, and his arguments proved influential. Several US economists embraced and built on them, and during the 1920s and 1930s, popular statistics textbooks and essays on statistical methodology quoted Keynes in passages dismissing the usefulness of the probability-theory-based measures (Biddle 2017, this volume; Lenel, this volume). A common theme in this American literature was that time series data, the type of data most often used in empirical research in economics in the first half of the twentieth century, did not fit the assumptions on which the inferential measures derived from probability theory were based. It was implausible, so the argument went, to regard a time series as a random sample from a larger universe characterized by stable relationships between variables. Further, even if one were willing to accept this characterization of a time series, the individual observations in a single time series sample were almost never independent of one another, a situation that contradicted one of the key assumptions underlying the probabilistic inferential methods of the time (Klein 1997). The rejection of the assumptions necessary for application of those inferential methods to time series data necessitated the development of alternative methods.

Along with his rejection of probability theory, Keynes offered an alternative framework for thinking about statistical inference. His central

theme was that the logic underlying good statistical induction (a phrase Keynes considered synonymous with statistical inference) was similar to the more familiar logic of universal induction, that is, the process of reasoning on the basis of multiple instances of observation to formulate and build confidence in conclusions claiming universality, for example, "all swans are white." Both forms of induction relied on what Keynes called the method of analogy. Inductive arguments in support of universal statements were built on numerous instances of observation. The characteristics shared by a set of instances constituted a *positive analogy* of the set (each involved a swan, in each the swan was white), while differences in characteristics across the instances constituted a *negative analogy* (swans of different sizes, observed in different seasons, on different continents).

Keynes argued that it was through careful consideration of the positive and negative analogy in sets of observations that one refined and/or built confidence in inductive generalizations. An additional observation of a white swan on a new continent would, in Keynes's terms, "strengthen the negative analogy" of one's set of instances and strengthen confidence in the universality of the positive analogy "all swans are white"; observing a black swan in Australia would narrow the scope of the generalization that the set of observations could support ("all swans outside Australia are white"), and so forth. The ultimate goal of Keynes's discussion, however, was to show how the same logic could be applied to the problem of statistical inference, that is, reasoning from samples of statistical data to probabilistic generalizations about events or relationships of the form "if A, then a 20 percent chance of B."

In statistical inference, one built general conclusions not on sets of individual observational instances but on sets of samples of statistical data. Each sample was like previous samples in some ways (the positive analogy) and unlike those samples in other ways. Further, any given sample was being used to draw conclusions about some "universe," with which it would share some characteristics but from which it would differ in certain ways. Building strong inferences on the basis of statistical measures required careful attention to the circumstances surrounding the generation of the data used to calculate the measures and the circumstances surrounding the phenomena about which one wished to draw conclusions. Keynes illustrated this point with the example of drawing an inference about the relationship between age and the probability of death on the basis of a sample of deceased individuals:

> We note the proportion who die at each age, and plot a diagram which displays these facts graphically. We then determine by some method of curve fitting a mathematical frequency curve which passes with close approximation through the points of our diagram. . . . In providing this comprehensive description the statistician has fulfilled his first function. But in determining the accuracy with which this frequency curve can be employed to determine the probability of death at a given age in the population at large, he must pay attention to a new class of considerations and must display a different kind of capacity. He must take account of whatever extraneous knowledge may be available regarding the sample of the population which came under observation, and of the mode and conditions of the observations themselves. Much of this may be of a vague kind, and most of it will be necessarily incapable of exact, numerical, or statistical treatment. (Keynes [1921] 1973: 372)

That Keynes's rejection of the inferential measures derived from probability theory was consistent with his positive heuristics for statistical inference can be seen in the quoted passage because the measures he rejected claimed to be able to provide reliable conclusions about phenomena outside the sample based only on information from the sample itself—without considering "extraneous knowledge" regarding the sample and considering only those characteristics of the sample members amenable to mathematical treatment. Keynes understood the arguments from probability theory that justified the use of these inferential measures, but he believed that the assumptions on which those arguments were based were seldom met in data from the social world (e.g., 418–19). Before one could arrive at an inference, one needed to ascertain the commonalities in the data that created a positive analogy, which Keynes in his *Treatise* called "uniformity" and in his controversy with Tinbergen on the econometric method, "homogeneity" (Boumans 2019). This aspect also played a decisive role on the justification of Galton's pictorial inference (Boumans, this volume): inferences from samples whose members did not have something in common—for example, belonging to the same "natural class"—would be meaningless. In biology, when drawing inferences from observations of the same species, for example, the requirement of uniformity could be justified. But with respect to economic or social phenomena this was an open question that first needed to be investigated (Klein 1997).

As with the rejection of probability-based inference, one finds in the US literature that explicitly addresses appropriate empirical methods echoes

of Keynes's constructive advice on *statistical induction*, sometimes accompanied by references to his *Treatise*. Biddle (this volume) shows how the inferential practices of an important group of empirical economists, those employed by the US Department of Agriculture in the 1920s and early 1930s, exemplified Keynes's ideas and admonitions about statistical inference.

More generally, during the prewar period economists employed a wide variety of strategies for developing methods for the statistical estimation of such things as the course of the price level, the effectiveness of various farming practices, and the relationship between the cyclical movements of different economic activities. In the early 1920s, the NBER was founded, and across the decades researchers associated with the NBER developed a distinctive methodology of empirical research, supplementing their publications with detailed descriptions of their inferences.[4] The NBER researchers were also instrumental in the development of the techniques of national income accounting, devising clever methods for estimating unknown quantities from observable data and assessing the reliability of those estimates. Many of these methods, which have little relationship to probability theory, remain part of national income accounting today. And, as Akhabbar (this volume) describes, Leontief created new methods of statistical inference for use with his interindustry or input-output analyses.

In the early 1940s, in the introduction to his "Probability Approach in Econometrics," Haavelmo (1944: iii) acknowledged, and promised to refute, the prevalent opinion among empirical economists that applying probability models was a "crime in economic research" and "a violation of the very nature of economic data." At the center of his refutation was an ingenious reconceptualization of the inferential problem presented by a sample of statistical data. In response to the doubts expressed by the leading statistical economists of the 1920s and 1930s about the wisdom of regarding a time series as a sample drawn from some known, fixed *universe*, Haavelmo proposed the idea of the time series as a set of observations generated by a mechanism, one capable of generating an infinity of observations. The mechanism could be characterized by a probability law, and the task of statistical inference was to discover that probability law (iii, 48). Haavelmo's reconceptualization of the economists' inferential problem was embraced by the Cowles Commission econometricians of

4. Stapleford's contribution to this volume revisits the ideas about the proper approach to statistical analysis and the role of that analysis in economics espoused and modeled by the NBER's intellectual leader, Mitchell.

the 1940s and early 1950s, who employed an array of established and new inferential procedures derived from probability theory.

This universe created by Haavelmo implied, however, a specific ontology: it was Nature's "enormous laboratory" that produced a "stream of experiments" (14). In this universe, samples are the outcomes of repeated experiments. This is a different universe than that of Keynes, which consists of people's beliefs and expectations and implies a different concept of probability. By putting statistical inference in an experimental setting, it created, according to Leamer (1983), the myth of empirical research being objective and free of personal prejudice.

After World War II, inspired by Keynes's *General Theory* and aided by newly developed methodologies, including national income accounting, macro-econometric modeling, and input-output analysis, economic policymaking was increasingly based on statistical analyses. One sees both probabilistic and nonprobabilistic inferential methods being employed in this work. Data-based approaches to economic policy development and implementation were being designed and further developed at national levels but also at newly founded international organizations like the United Nations, the World Bank, and the International Monetary Fund. There are of course institutional histories of these organizations, but they give relatively little attention to the statistical approaches on which they based their policy programs (but see Samuel, this volume, and references cited therein).

The early 1960s mark something of a turning point in the pedagogy of statistics and econometrics, after which graduate students in economics would routinely be taught to understand statistical estimation and inference as an application of probability theory, whether in the context of a Cowles-style presentation of simultaneity, identification, and so forth, or simply more prosaic instruction in constructing confidence intervals and testing ordinary least squares regression coefficients for *statistical significance*. In the 1950s, however, the amount of systematic instruction in methods of statistical inference based in probability theory available to interested economics graduate students, not to say average economics graduate students, depended on where they were being trained. For example, it was not until 1962 that a departmental committee at Columbia University, home of one of the largest economics PhD programs in the United States, recommended that econometrics be offered as a field for graduate students (Rutherford 2004).

At the University of Chicago, site of another major US graduate program and home of the Cowles Commission until 1955, the situation was

complicated. Griliches came to Chicago in 1954, took classes from Henri Theil and Haavelmo, among others, and began teaching graduate econometrics there himself in 1957. Among his students in the early 1960s was the future econometrician G. S. Maddala, who considered writing a dissertation in econometric theory before opting for a more empirically oriented topic. But Maddala reports that econometrics at Chicago was very "low key." Neither Maddala nor any other student in his cohort who was writing an empirical dissertation used anything more complicated than ordinary least squares regression. High-tech methods were eschewed for actual empirical work (Krueger and Taylor 2000; Lahiri 1999). So, what did inference look like in these empirical dissertations and in the work of the Chicago economists who supervised them?

Burnett (this volume) provides a good account of the inferential method employed by one leading University of Chicago economist, Theodore Schultz, and in the dissertation of Griliches, one of Schultz's most successful students. Schultz reported standard errors and formal hypothesis tests, but they were almost irrelevant to his arguments. The same can be said of the inferential arguments found in the work of other influential contributors to the "Chicago Economics" of that era, including Milton Friedman, H. Gregg Lewis, and Jacob Mincer (Biddle 2017). More research that took a close look at the techniques of statistical inference taught and employed at the University of Chicago during this period would be welcome.

Just as it took time for the inferential methods associated with Cowles-style econometrics to dominate economic pedagogy, it took time for those methods to spread through the empirical literature in economics. In the 1940s and 1950s, most empirical articles in economics did not use regression methods, much less report standard errors or tests of statistical significance (Backhouse 1998; Biddle 1999). By the 1960s, regression analysis had become the preferred method of detecting and measuring the economic relationships involved in theoretical analyses, but a number of influential empirical studies that used regression analysis made little use of inferential techniques derived from probability theory. This is due partly, no doubt, to simple inertia—even by 1960, empirical economics was dominated by people who had little or no formal training in the use of the Cowles-inspired inferential methods and did not see a need to learn them. But, as Biddle (2017) argues, neither Haavelmo's essay nor the empirical methods used by the Cowles Commission econometricians provided convincing answers for several important elements of the preexisting case

against applying probability theory to economic data, and many empirical economists of the 1940s and 1950s may have carefully considered the increasingly popular inferential procedures and consciously decided against using them. Kuznets (1950), Leontief, and Millard Hastay (1951) were among the prominent empirical economists who explicitly expressed doubts about the value of the new methods (Akhabbar, this volume).

Nonetheless, by the 1970s, there was a broad consensus in the profession that inferential methods justified by probability theory—methods of producing estimates, of assessing the reliability of those estimates, and of testing hypotheses—were not only applicable to economic data but a necessary part of almost any attempt to generalize on the basis of economic data. In discussing the nature of this consensus, and how it differed from the reigning opinions on statistical inference held by the empirical economists of thirty years earlier, it is helpful to make use of the concept of mechanical objectivity introduced by Lorraine J. Daston and Peter Galison (1992) in their writings on the history of scientific objectivity and fruitfully applied to the history of quantification in the social sciences by Theodore Porter in his 1995 book *Trust in Numbers*.

Statistical inference is about using samples of statistical data as a basis for drawing conclusions about what is true, or probably true, in the world beyond the sample. In this setting, mechanical objectivity means employing a set of explicit and detailed rules and procedures to produce conclusions that are objective in the sense that if many different people took the same statistical information, and followed the same rules, they would come to exactly the same conclusions. The trustworthiness of the conclusion depends on the quality of the method. *Statistical inference* as defined and described in post-1960 econometrics textbooks is a prime example of this sort of mechanical objectivity.

Porter contrasts mechanical objectivity with an objectivity based on the "expert judgment" of those who analyze sample data. The analyst's expertise is acquired through a training process sanctioned by a scientific discipline, as well as through experience making similar decisions using similar data subject to the surveillance of other experts. One's faith in the analyst's conclusions depends largely on one's assessment of the quality of his or her disciplinary expertise but also on his or her commitment to the ideal of scientific objectivity.

Speaking in these terms, we would argue that in the 1920s and 1930s, the importance and propriety of applying expert judgment in the process of statistical inference was explicitly acknowledged by empirical econo-

mists. At the same time, mechanical objectivity was valued—it is easy to find examples of empirical economists employing rule-oriented, replicable procedures for drawing conclusions from economic data. The rejection of the tools of inference based on probability theory during this period was simply a rejection of one particular technology for achieving mechanical objectivity. In the post-1970s consensus regarding statistical inference in economics, however, application of this one particular form of mechanical objectivity became an almost required part of the process of drawing conclusions from economic data, taught in a standardized way to every economics graduate student.

Also, although there is a fundamental tension between the desire for mechanically objective methods and the belief in the importance of expert judgment in arriving at and communicating statistical results, it would be wrong to characterize what happened to statistical inference between the 1940s and the 1970s as a displacement of procedures requiring expert judgment by mechanically objective procedures. In the 1920s and 1930s there was disagreement over whether the phrase *statistical inference* should be applied to all aspects of the process of drawing conclusions based on statistical data, or whether it meant only the use of formulas derived from probability theory to create estimates from statistical data, measure the reliability of those estimates, and use those estimates to test hypotheses. The econometrics textbooks published after 1960 explicitly or implicitly accepted this second, narrower, definition, and their instruction on statistical inference was largely limited to instruction in the mechanically objective procedures based on probability theory. It was understood, however, that expert judgment was still an important part of empirical economic analysis, particularly in the specification of the economic models to be estimated. But the disciplinary knowledge needed for this task was to be taught in other classes, using other textbooks.

And something else was left largely unspoken in the descriptions of procedures for statistical inference found in the econometric textbooks from this period: even after choosing the statistical model, calculating the estimates and standard errors, and conducting the hypothesis tests, there was room for an empirical economist to exercise a fair amount of judgment, based on his or her specialized knowledge, before drawing conclusions from the statistical results. Indeed, no procedure for drawing conclusions from data, no matter how algorithmic or rule bound, can dispense entirely with the need for expert judgment (Boumans 2015: 84–85). And few empirical economists after 1970 would deny that the interpretation of

statistical results, even those produced and assessed using the methods that had come to be called "econometrics," often involved a good deal of expert judgment. These "broader" inferential skills, such as those endorsed by Keynes, however, became a sort of craft knowledge picked up by economists from thesis advisers or other mentors, often referred to as an "art," thereby separating it from science itself.

This does not mean that the near ubiquity of a set of standard inferential tests and measures associated with probability theory in the empirical economics literature since the 1970s was simply a change in style or rhetoric. When application of these inferential procedures became a necessary part of economists' analyses of statistical data, the results of applying those procedures came to act as constraints on the set of claims that a researcher could credibly make to his or her peers on the basis of that data. For example, if a regression analysis of sample data yielded a large and positive partial correlation, but the correlation was not *statistically significant*, it would simply not be accepted as evidence that the *population* correlation was positive. If estimation of a statistical model produced a significant estimate of a relationship between two variables, but a statistical test led to rejection of an assumption required for the model to produce unbiased estimates, the evidence of a relationship would be heavily discounted. And once an author had justified an empirical model with a theoretical argument, a presentation and discussion of the coefficient estimates and significance tests associated with one or two versions of the empirical model was often a sufficient amount of *interpretation of results* to satisfy journal editors and referees.

So, we believe that in the latter half of the twentieth century, a mechanically objective procedure to generalize on the basis of statistical measures went from being a choice determined by the preferences of the analyst to a professional requirement, one that had real consequences for what economists would and would not assert on the basis of a body of statistical evidence. At the same time, the results produced by the procedure were still only part of the argument, to be combined with theory and other forms of evidence. This raises the question of whether one can discern implicit, but widely observed, *canons of inference* in the post-1970 empirical literature of this period. Arguably, the *credibility revolution* in the empirical microeconomic literature of the 1990s (Angrist and Pischke 2010), recently examined by Matthew T. Panhans and John D. Singleton (2017), represents among other things a significant change in these implicit canons of inference.

Stapleford (this volume) raises another interesting possibility, suggesting that the period during which the use of probability-based inferential tools is regarded as a necessary part of empirical research in economics may be an interlude in the history of economics that is coming to an end. Looking at the methodological statements and research practices of economists associated with the *data revolution* that commenced in the late twentieth century, he makes the case that the approach to empirical research being adopted by modern economists working with *big data* bears a resemblance to one promoted by Mitchell at the NBER, in which there was little place for tools and methods based in probability theory.

Finally, one still finds in the last decades of the twentieth century many examples of statistical inference without probability. Input-output models, computational general equilibrium models, and macroeconomic models with calibrated parameter values all represent techniques for using data to generate estimates of economic quantities and relationships. The growth accounting methods developed in the 1950s and 1960s to explain past economic growth, and applied to understand the productivity slowdown of the 1970s and 1980s, were seen to be largely *noneconometric* in nature (Biddle 2021: chap. 6). National income accounts continued to be updated, and international agencies like the IMF and the World Bank continued to create statistical pictures of national economies to guide their decisions (see Samuel, this volume). Economists working with all these empirical approaches faced the inferential problems of determining the reliability and the generalizability of the estimates they produced. For the most part, however, the nonprobabilistic inferential methods these researchers developed were not part of the formal econometrics curriculum of typical graduate programs in economics. Instead, they were taught in optional courses in places where an expert in the area might be part of the faculty, or picked up as the young PhDs found themselves in need of apprentice-like instruction from members of the established community of researchers working with those methods. Again, this situation has rendered these methods less visible to historians, and the question of how economists using these models did inference, and how their inferences were regarded by a profession that was for the most part committed to the probability-based approach, seems well worth pursuing.

References

Angrist, Joshua, and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24, no. 2: 3–30.

Backhouse, Roger. 1998. "The Transformation of US Economics, 1920–1960, Viewed through a Survey of Journal Articles." In *From Interwar Pluralism to Postwar Neoclassicism*, edited by Mary S. Morgan and Malcolm Rutherford. *History of Political Economy* 30 (supplement): 85–107.

Biddle, Jeff E. 1999. "Statistical Economics, 1900–1950." *History of Political Economy* 31, no. 4: 607–52.

Biddle, Jeff E. 2017. "Statistical Inference in Economics, 1920–1965: Changes in Meaning and Practice." *Journal of the History of Economic Thought* 39, no. 2: 149–74.

Biddle, Jeff E. 2021. *Progress through Regression: The Life History of the Empirical Cobb-Douglas Production Function*. Cambridge: Cambridge University Press.

Boumans, Marcel. 2015. *Science outside the Laboratory*. Oxford: Oxford University Press.

Boumans, Marcel. 2019. "Econometrics: The Keynes-Tinbergen Controversy." In *The Elgar Companion to John Maynard Keynes*, edited by Robert W. Dimand and Harald Hagemann, 283–89. Cheltenham, UK: Edward Elgar.

Boumans, Marcel, and Ariane Dupont-Kieffer. 2011. "A History of the Histories of Econometrics." In Boumans, Dupont-Kieffer, and Qin 2011: 5–31.

Boumans, Marcel, Ariane Dupont-Kieffer, and Duo Qin, eds. 2011. *Histories on Econometrics*. Supplemental issue to vol. 43 of *History of Political Economy*. Durham, NC: Duke University Press.

Bowley, Arthur. 1920. *Elements of Statistics*. 4th ed. London: P. S. King and Son.

Burns, Arthur F., and Wesley C. Mitchell. 1946. *Measuring Business Cycles*. New York: National Bureau of Economic Research.

Daston, Lorraine J., and Peter Galison. 1992. "The Image of Objectivity." In "Seeing Science." Special issue, *Representations* 40: 81–128.

De Marchi, Neil, and Christopher Gilbert, eds. 1989. *History and Methodology of Econometrics*. Oxford: Oxford University Press.

Edgeworth, F. Y. 1913. "On the Use of the Theory of Probabilities in Statistics Relating to Society." *Journal of the Royal Statistical Society* 76, no. 2: 165–93.

Epstein, R. 1987. *A History of Econometrics*. Amsterdam: North-Holland.

Greene, William H. 1999. *Econometric Analysis*, 4th ed. Englewood Cliffs, NJ: Prentice Hall.

Haavelmo, Trygve. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 (supplement): iii–115.

Hastay, M. 1951. Review of *Statistical Inference in Dynamic Economic Models,* edited by J. Marschak and T. Koopmans. *Journal of the American Statistical Association* 46, no. 255: 388–90.

Hoover, Kevin D., and Michael Dowell. 2001. "Measuring Causes: Episodes in the Quantitative Assessment of the Value of Money." In *The Age of Economic Measurement*, edited by Judy L. Klein and Mary S. Morgan. *History of Political Economy* 33 (supplement): 137–61.

Keynes, John Maynard. (1921) 1973. *A Treatise on Probability*. London: Macmillan.

Klein, Judy L. 1997. *Statistical Visions in Time: A History of Time Series Analysis, 1662–1938*. Cambridge: Cambridge University Press.

Klein, Judy L., and Mary S. Morgan, eds. 2001. *The Age of Economic Measurement*. Supplemental issue to vol. 33 of *History of Political Economy*. Durham, NC: Duke University Press.

Koopmans, Tjalling C. 1947. "Measurement without Theory." *Review of Economics and Statistics* 29, no. 3: 161–72.

Krueger, Alan B., and Timothy Taylor. 2000. "An Interview with Zvi Griliches." *Journal of Economic Perspectives* 14, no. 2: 171–89.

Kuznets, Simon. 1950. "Conditions of Statistical Research." *Journal of the American Statistical Association* 45, no. 249: 1–14.

Lahiri, K. 1999. "The ET Interview: G. S. Maddala." *Econometric Theory* 15, no. 5: 753–76.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*. New York: Wiley.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73, no. 1: 31–43.

Maas, Harro, and Mary S. Morgan, eds. 2012. *Observing the Economy*. Supplemental issue to vol. 44 of *History of Political Economy*. Durham, NC: Duke University Press.

Morgan, Mary S. 1990. *The History of Econometric Ideas*. Cambridge: Cambridge University Press.

Morgan, Mary S. 2012. *The World in the Model*: *How Economists Work and Think*. Cambridge: Cambridge University Press.

Panhans, Matthew T., and John D. Singleton. 2017. "The Empirical Economist's Toolkit: From Models to Methods." In *The Age of the Applied Economist: The Transformation of Economics since the 1970s*, edited by Roger E. Backhouse and Béatrice Cherrier. *History of Political Economy* 49 (supplement): 127–57.

Porter, Theodore. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Qin, Duo. 1993. *Formation of Econometrics: A Historical Perspective*. Oxford: Oxford University Press.

Qin, Duo. 2013. *A History of Econometrics: The Reformation from the 1970s*. Oxford: Oxford University Press.

Rutherford, Malcolm. 2004. "Institutional Economics at Columbia University." *History of Political Economy* 36, no. 1: 31–78.

Schultz, Theodore W. 1964. *Transforming Traditional Agriculture*. New Haven, CT: Yale University Press.

Woody, Andrea I. 2014. "Chemistry's Periodic Law: Rethinking Representation and Explanation after the Turn to Practice." In *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science*, edited by Léna Soler, Sjoerd Zwart, Michael Lynch, and Vincent Israel-Jost, 123–50. New York: Routledge.

Yule, G. U. 1911. *Introduction to the Theory of Statistics*. London: Charles Griffen.