IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Mathematics and Statistics in the Social Sciences

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1662639 since 2020-05-06T15:39:43Z

Publisher:

SAGE

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

12 October 2024

Mathematics and Statistics in the Social Sciences

Stephan Hartmann*and Jan Sprenger†

May 7, 2010

Over the years, mathematics and statistics have become increasingly important in the social sciences[1]. A look at the history quickly confirms this claim. At the beginning of the 20th century most theories in the social sciences were formulated in qualitative terms while quantitative methods did not play a substantial role in the formulation and establishment of them. Moreover, many practitioners considered mathematical methods to be inappropriate and simply not suited to foster our understanding of the social domain. Notably, the famous *Methodenstreit* was also about the role of mathematics in the social sciences. Here mathematics was considered to be the method of the natural sciences from which the social sciences had to be separated during the period of maturation of these disciplines. All this changed by the end of the century. By then, mathematical and especially statistical methods were standardly used and it became relatively uncontested that they are of much value in the social sciences. In fact, the use of mathematical and statistical methods is now ubiquitous: Almost all social sciences rely on statistical methods to analyze data and to form hypotheses, and almost all of them use (to a greater or lesser extend) a range of mathematical methods to help us understand the social world.

Additional indication for the increasing importance of mathematical and statistical methods in the social sciences is the formation of new sub-disciplines, and the establishment of specialized journals and societies. And indeed, sub-disciplines such as Mathematical Psychology and Mathematical Sociology emerged, and corresponding journals such as *The Journal of Mathematical Psychology* (since 1964), *The Journal of Mathematical Sociology* (since 1976), *Mathematical Social Sciences* (since 1980) as well as the online journals *Journal of Artificial Societies and Social Simulation* (since 1998) and *Mathematical Anthropology and Cultural Theory* (since 2000) were established. What is more, societies, such such as the Society for Mathematical Psychology (since 1976) and the Mathematical Sociology Section of the American Sociological Association (since 1996) were founded. Similar developments happened in other countries.

The mathematization of economics set in somewhat earlier (Vazquez 1995; Weintraub 2002). However, the use of mathematical methods in economics started booming only in the second half of the last century (Debreu 1991). Contemporary economics is dominated by the mathematical approach, although a

*Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: s.hartmann@uvt.nl, webpage: www.stephanhartmann.org.

†Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, e-mail: j.sprenger@uvt.nl, webpage: www.laeuferpaar.de.

[1]In our usage, "social science" includes disciplines such as anthropology, political science, and sociology, but also economics and parts of linguistics and psychology.

certain *style* of doing economics becomes more and more under attack in the last decade or so. Recent developments in behavioral economics and experimental economics can also be understood as a reaction against the dominance (and limitations) of an overly mathematical approach to economics. There are similar debates in other social sciences, but it is important to stress that problems of one method (such as axiomatization or the use of set theory) can hardly be taken as a sign of bankruptcy of mathematical methods in the social sciences *tout court*.

This chapter surveys mathematical and statistical methods used in the social sciences and discusses some of their philosophical questions. It is divided into two parts. Sections 1 to 3 are devoted to mathematical methods, and Sections 4 to 8 to statistical methods.

1 A Plurality of Mathematical Methods

Social scientists use a wide variety of mathematical methods. Given that the space of this chapter is restricted, it is impossible to list them all, give examples, examine their domain of applicability, and to discuss some of the philosophical problems they raise. Instead, we broadly distinguish between three different kinds of methods: (i) methods imported from the formal sciences, (ii) methods imported from the natural sciences, and (iii) *sui generis* social scientific methods. We review them in turn.

Methods imported from the formal sciences include (linear) algebra, calculus (including differential equations), the axiomatic method, logic and set theory, probability theory (including Markov chains), linear programming, topology, graph theory, and complexity theory. All these methods have important applications in the social sciences.[2] In recent years, various methods from computer science have been incorporated into social science research as well. There is also a strong trend within computer science to address problems from the social sciences. An example is the recent establishment of the new interdisciplinary field *Computational Social Choice* which is dominated by computer scientists.[3]

Interesting, though, much work in Computational Social Choice uses analytical and logical methods. There is, however, also a strong trend in the social sciences to use powerful numerical and simulation methods to explore complex an (typically) dynamical social phenomena. The reason for this is of course the availability of high-powered computers. But not all social scientists follow this trend. Especially many economists are reluctant to use simulation methods and do not consider them to be appropriate tools to study economic systems.[4]

Methods imported from the natural sciences become more and more popular in the social sciences. These methods are more specific than the formal methods mentioned above; they involve substantial assumptions that happen – or so it is claimed – to be fulfilled in the social domain. These methods comprise tools to study multi-agent systems, the theory of complex systems, non-linear

[2]For a lucid exposition of many of these methods and examples from the social sciecnes, see Luce and Suppes (1968).

[3]See http://www.illc.uva.nl/COMSOC/

[4]For a discussion of computer simulations in the social sciences, see Hegselmann et al. (1996). In this context it is interesting to study the influence of the work done at the Santa Fe Institute on mainstream economics. See e.g. Anderson et al (1988). See also Waldrop (1992).

dynamics, and the methods developed in synergetics (Weidlich 2006) and, more recently, in econophysics (Mantegna and Stanley 1999). The applicability of these methods follows from the observation that societies are nothing but many-body systems (like a gas is a many-body system composed of molecules) that exhibit certain features (such as the emergence of ordering phenomena). Hence, these features can be accounted for in terms of a statistical description, just like gases and other many-body systems which are studied in the natural sciences. Such methods are also used in new interdisciplinary fields such as environmental economics.

Besides providing various methods that can be used to study social phenomena, the natural sciences also inspired a certain way of addressing a problem. Meanwhile it is common to say that the core activity in the social sciences is *model building*.[5] These models are carefully crafted, idealizations have to be made, and the consequences of the model are often obtained with the help of a *computer simulation*. This is much like in physics. Let us therefore call this approach the physicist's approach to social science and contrast it with the mathematician's approach to social science outlined above.

Finally, there are **mathematical methods that emerged from problems in the social sciences**. These include powerful instruments such as decision theory[6], utility theory, game theory[7], measurement theory (Krantz et al. 1971), social choice theory (Gaertner 2006), and Judgment Aggregation (List and Puppe 2009). The latter theories were invented by social scientists, for social scientists and with a specific social-science application in mind. They help addressing specific problems that arise in the context of the social sciences that did not have an analogue in the natural sciences when they were invented. Only later, some of these theories also turned out to be useful in the natural sciences or have been combined with insights from the natural sciences. Evolutionary game theory is a case in point.[8] Other interesting examples are the study quantum games (Piotrowski and Sladkowski 2003) and the application of decision theory in fundamental physics (Wallace 2010).

Interestingly, there are also methods that cannot be attached to one specific science. Network theory is a case in point. As networks are studied in almost all sciences, parallel developments took place, and much can be learned by exploring what's been obtained in other fields (Jackson 2008).[9]

Having listed a large number of methods, the question arrises which method is appropriate for a certain problem. This question can only be answered on a case by case basis and it is part of the ingenuity of the scientist to pick the best method. But let us stress the following: While some scientists ask themselves which problems they can address with their favorite method, the starting point should always be a specific problem. And once a problem is chosen, the scientist picks the best method that helps solving it. To have some choice, it is important that scientists are acquainted with as many methods as possible. Mathematics (and related disciplines) provide the scientist with a toolbox out of which they

[5]For a more detailed discussion of modeling in the social sciences, see ch. 29 ("Local models versus global theories, and their assessment") of this handbook. For a general review of models in science, see Frigg and Hartmann (2007).

[6]See ch. 15 ("Rational choice and its alternatives") of this handbook.

[7]See ch. 16 ("Game theory") of this handbook.

[8]See ch. 17 ("Evolutionary approaches") of this handbook.

[9]See also ch. 18 ("Networks") of this handbook.

have to pick a tool. Let us call this the *Toolbox View*.

2 Why Mathematizing the Social Sciences?

A historically important reason for the mathematization of the social sciences was that 'mathematics' is associated with precision and objectivity. These are (arguably) two requirements any science should satisfy, and so the mathematization of the social sciences was considered to be a crucial step that had to be taken to turn the social sciences into real science. Some such view has been defended by many authors. Luce and Suppes (1968), for example, argue along these lines for the importance of axiomatizing the theories of the social sciences. These authors also developed measurement theory (Krantz et al. 1971) and Suppes (1967, 2001) showed how the relation between a theory and its target system can be explicated in mathematical terms. Contrary to this tradition, it has been argued that the subject matter of the social sciences does not require a high level of precision and that the social sciences are and should rather be inexact (cf. Hausman 1992). After all, what works in the natural sciences may well not work in the social sciences.

While Sir Karl Popper, one of the towering figures in the methodology of social science, did not promote the mathematization of the social sciences in the first place (Hands 2008), it is clear that it nevertheless plays an enormous role in his philosophy. Given his focus on predictions and falsifiability, a theory that is mathematized is preferable to a theory that is not. After all, it is much easier to derive falsifiable conclusions from clearly stated propositions than from vague and informal claims.

It is a mistake, however, to overestimate the role of the mathematics. At the end, mathematics provides the social scientist only with tools, and what one obtains when using these tools will crucially depend on the assumptions that are made. This is a variant of the well known GIGO principle from computer science ("garbage in, garbage out). All assumptions are motivated informally; formulating them in the language of mathematics just helps putting them more precisely. And once the assumptions are formulated mathematically, the machinery of mathematics helps to draw inferences in an automated way. This holds for analytical calculations as well as for numerical studies, including computer simulations (Frigg and Reiss 2010; Hartmann 1996).

This brings us to another advantage of mathematical methods in the social sciences. While non-formal theories often remain rather simplistic and highly idealized, formal theories can be complicated and (more) realistic, reflecting the messiness of our world. The mathematical machinery then helps drawing inferences which could not be obtained without them (Humphreys 2004). Often different assumptions pull in opposite directions, and it is not clear which one will be stronger in a specific situation. However, when implemented in a mathematical model, it can be derived what happens in which part of the parameter space. And so the availability of powerful computers allows the systematic study of more realistic models.

There is, however, also a danger associated with this apparent advantage. Given the availability of powerful computers, scientists may be tempted to construct very complex models. But while these models may do well in terms of empirical adequacy, it is not so clear that they also provide *understanding*. This

is often provided by rather simple models (sometimes called 'toy models'), i.e. models that pick only one crucial aspect of a system and help us to get a feel for what follows from it.[10]

There are several other reasons for the mathematization in the social sciences. We list them in turn.

1. **Theory Representation.** Mathematics is used to formulate a theory. By doing so, the structure of the theory becomes transparent and the relationships that hold between the variables can be determined. Mathematics provides clarity, generality, and rigor. There are many ways to represent a theory. For long, philosophers have championed the syntactic view (basically a representation of the theory in first order logic) or the semantic view in its various forms (Balzer et al 1987; Suppes 2000). While these reconstructions may be helpful for coming up with a consistent version of a theory, it apparently suffices for all practical purposes to state a set of equations that constitute the mathematical part of the theory.

2. **Theory Exploration.** Once the theory is represented in mathematical terms, the mathematical machinery can be employed to derive qualitative and quantitative consequences of the theory. This helps to better understand what the theory is all about and what it entails about the world. The deductive consequences of the theory (and additional assumptions that have to be made) can be divided into retrodictions or predictions. For retrodictions the question arises which additional assumptions have to be made to obtain a certain (already measured) value of a variable.

3. **Theory Testing.** The predictions of a mathematically formulated theory can then be used to test the theory by confronting its consequences with relevant data. At the end, the theory will be confirmed or disconfirmed, or to put in Popperian terms, corroborated or falsified.

4. **Heuristics.** Once the structure of a theory is formulated in mathematical term, a look at it may reveal analogies to other phenomena. This may inspire additional investigations ("intuition pump") and lead to a better understanding of the class of phenomena under investigation. Also, a numerical study of a theory may suggest new patterns that can be incorporated into the assumptions of another theory. A good example of this point is the use of cellular automata for studying the emergence of ordering phenomena, such as in Schelling's famous Segregation Model (Sugden 2008).

5. **Explanation and Understanding:** While it is controversial what a scientific explanation is, it is clear that – once the theory is formulated mathematically – a phenomenon can be fitting into a larger theoretical pattern (as the unification account demands) or a causal story can be read off from the theory.

[10]For more doubts about some of the uses of simulations in the social sciences, see Humphreys (2004).

3 Methodological Issues

There are interesting parallels between the use of mathematics in the natural and social sciences. In both kinds of sciences, we find a plurality of methods ranging from axiomatic methods to the use of computer simulations. We also find very different types of models, ranging from toy models (that illuminate one feature of a system in a simple way without scoring high in terms of empirical adequacy) to models that fit a large amount of data (but do not provide much understanding). The mathematization also has similar purposes in both kinds of sciences: it helps to represent a certain object or system, to explain it and to make predictions to test the underlying theory or model.

However, there is also an interesting difference. This difference has to do with the relation between the mathematical formalism and the data in the natural and social sciences. Let us assume that we have constructed a mathematical model and we confront it with data. If the data correspond to what the model predicts, the model is confirmed. If the data contradict the model's prediction, then there are two options in the natural sciences: either there was a measurement error, or (or an error can be excluded) the model has a problem. These two possibilities also show up in many social-science contexts. However, there is a third option in the social sciences which has to do with the observation that the data in the social sciences are often not very hard.

Let us give an example from cognitive psychology. In a series of experiments, Tversky and Kahneman (1983) showed that the participants commit fallacies such as the conjunction fallacy. In that case, 85% of the participants judge the conjunction of two propositions to be more probable than on the the conjuncts. One option to argue is that these people are irrational as they violate a basic rules of probability. However, things are not that easy. In the sequel, many other experiments were conducted and proposals were made as to how people reason in such cases. What results is an intricate interplay between mathematical modeling and experimentation which does not occur in this form in the natural sciences (Hertwig et al 2008; Hartmann and Meijs 2010).

The softness of the data is probably also one of the reasons why there is much more debate in the social sciences about the usefulness of mathematical methods. The social sciences exhibit a wealth of of different approaches, and mathematical methods play a more or less important role in them. The defenders of mathematical methods will argue that mathematics simply provides a host of structures, and as the social world is structured just like the natural world is, some of these structures will fit (or approximately fit). Opponents will either doubt that there are stable structures to be found in the social world, or they will argue that the structures that mathematics (and related sciences) provide do not fit as the social world is very different from the natural world. We take this to be an empirical question and do not see a reason why one should not examine go ahead and employ mathematics in the social sciences.

Besides the general debate about the usefulness of mathematical methods in the social sciences, there is also a lot of debate about the question which methods are most appropriate. An example is the debate about econophysics. The practitioners of this field (mostly physicists or social scientists with a background in physics) approach certain problems from economics with the tools of statistical physics. Typical results are the explanation of certain power laws that show up in economical data. But while concepts and ideas from physics

have played an important role in economics in the past[11], many economists do not consider the explanations given by econophysicists as appropriate, cf. Gallegatti (2006).

The above mentioned example from cognitive psychology suggests, however, that the standards for the assessment of mathematical approaches to the social sciences are much less clear than in the natural sciences. One may reasonably doubt the use of mathematical models altogether, and one may also doubt the application of a specific method. Much more needs to be done here until a consensus (if there will ever be one) is reached.

4 The Development of Statistical Reasoning

Statistical reasoning is nowadays a central method of the social sciences. First, it is indispensable for *evaluating experimental data* e.g. in behavioral economics or experimental psychology. For instance, psychologists might want to find out whether men act, in a certain situation, differently from women, or whether there are causal relationships between violent video games and aggressive behavior. Second, the social sciences heavily use statistical models as a *modeling tool* for analyzing empirical data and predicting future events, especially in econometrics and operational research, but recently, also in the mathematical branches of psychology, sociology, and the like. For example, time series and regression models relate a number of input (potential predictor) variables to output (predicted) variables. Sophisticated model comparison procedures try to elicit the structure of the data-generating process, eliminate some variables from the model, select a "best model" and finally fit the parameter values to the data.

Still, the conception of statistics as an *inferential* tool is quite young: throughout the 19th century, statistics was mainly used as a *descriptive* tool to summarize data and to fit models. While in inferential statistics, the focus lies on testing scientific hypotheses against each other, or quantifying evidence for or against a certain hypothesis, descriptive statistics focuses on summarizing data and fitting the parameters of a given model to a set of data. The most famous example is maybe Gauß' method of the least squares, a procedure to center a data set $(x_n, y_n)_{n\in\mathbb{N}}$ around a straight line. Other important descriptive statistics are contingency tables, effect sizes, and tendency and dispersion measures.

Descriptive statistics were, however, "statistics without probability" (Morgan 1987), or as one might also say, statistics without uncertainty. In the late 19th and early 20th century, science was believed to be concerned with *certainty*, with the discovery of invariable, universal laws. This left no place for uncertain reasoning. Recall that at that time, stochastic theories in the natural sciences, such as statistical mechanics, quantum physics, or laws of inheritance, were still quite new or not yet invented. Furthermore, there was a hope of reducing them to more fundamental, deterministic regularities, e.g. to take the stochastic nature of statistical mechanics as an expression of our imperfect knowledge, our uncertainty, and not as the fundamental regularities that govern the motion of molecules. Thus, statistical modeling contradicted

[11] Examples are the work of the Physiocrats and the introduction of the concept of equilibrium, see Mirrowki (1889).

the *nomothetic ideal* (Gigerenzer 1987), inspired by Newtonian and Laplacean physics, of establishing universal laws. Therefore statistics was considered as a mere auxiliary, imperfect device, a mere surrogate for proof by deduction or experiment. For instance, the famous analysis of variance (ANOVA) obtained its justification in the nomothetic view through its role in causal inference and elucidating causal laws.

It is interesting to note that these views were held even in the social sciences, although the latter dealt with a reality that was usually too complex to isolate causal factors in laboratory experiments. Controlling for external impacts and confounders poses special problems to the social sciences, whose domain are humans and not inanimate objects. The search for deterministic, universal laws in the social sciences might thus seem futile – and this is probably the received view today –, but in the first half of the 20th century many social scientists thought differently. Statistics was needed to account for measurement errors and omitted causal influences in a model, but it was thought to play a merely provisional role:

> "statistical devices are to be valued according to their efficacy in enabling us to lay bare the true relationship between the phenomena under consideration. An ideal method would eliminate all of the disturbing factors." (Schultz 1928, 33)

Thus, the view of statistics was *eliminativist*: as soon as it has done the job and elucidated the laws which we aim at, we can dismiss it. In other words, the research project consisted in eliminating probabilistic elements, instead of discovering statistical laws and regularities or modeling physical quantities as probabilistic variables with a certain distribution. This methodological presumption, taken from 19th century physics, continued to haunt social sciences far into the first half of the 20th century. Economics, as the "physics of social sciences", was particularly affected by that conception (Morgan 2002).

In total, there are three main reasons why inferential statistics was recognized as a central method of the social sciences:

1. The advances in mathematical probability, as summarized in the seminal work of Kolmogorov (1933/56).

2. The inferential character of many scientific questions, e.g. whether there is a causal relationship between variables X and Y. There was a need for techniques of data analysis that ended up with an inference or a decision, rather than with a description of a correlation.

3. The groundbreaking works by particular pioneer minds, such as Tinbergen and Haavelmo in economics (Morgan 1987).

The following sections investigate the different ways inferential statistics has been spelled out, with a focus on the school that is most prominent in modern social science: Fisher's method of significance testing.

5 Significance Tests and Statistical Decision Rules

One of the great conceptual inventions of the founding fathers of inferential statistics was the *sampling distribution* (e.g. Fisher 1935). In the traditional approach (e.g. in classical regression), there was no need for the concept of a sample drawn from a larger population – instead, the modeling process directly linked the observed data to a probabilistic model. In the modern understanding, the actual data are just a sample drawn out of a much larger, hypothetical population about which we want to make an inference. The rationale for this view of data consists in the idea that scientific results need to be *replicable*. Therefore, we have to make an inference about the comprehensive population (or the data-generating process, for that matter) instead of making an 'in-sample' inference whose validity is restricted to the particular data we observed. This idea of a sampling distribution proved crucial for what is known today as frequentist statistics. That approach strongly relies about this idea of the sampling distribution, outlined in the seminal works of Fisher (1925, 1935, 1956) and Neyman and Pearson (1933, 1967), parting ways with the classical accounts of Bayes, Laplace, Venn and others.

In frequentist statistics, there is a sharp division between approaches that focus on inductive *behavior*, such as the Neyman-Pearson school, and those that focus on inductive *inference*, such as Fisherian statistics. To elucidate the difference, we will present both approaches in a nutshell. Neyman and Pearson (1933) developed a behavioral framework for deciding between two competing hypotheses. For instance, take the hypothesis H_0 that a certain learning device does not improve the students' performance, and compare it to the hypothesis H_1 that there is such an effect. The outcome of the test is interpreted as a judgment on the hypothesis, or the prescription to take a certain action ("accept/reject H_0"). They contrast two hypotheses H_0 and H_1 and develop testing procedures such that the probability of erroneously rejecting H_0 in favor of H_1 is bounded at a certain level α, and that the probability of erroneously rejecting H_1 in favor of H_0 is, given that constraint, as low as possible. In other words, Neyman and Pearson aim at maximizing *power* of a test (the chance of a correct decision for H_1) under the condition that the *level* of the test (the chance of an incorrect decision for H_1) is bounded at a real number α. Thus, they developed a more or less symmetric framework for making a decision between competing hypothesis, with the aim of minimizing the chance of a wrong decision.

While such testing procedures apply well to issues of quality control in industrial manufacturing and the like, the famous biologist and statistician Ronald A. Fisher (1935, 1956) argued that they are not suitable for the use in *science*. First, a proper behavioristic, or decision-theoretic approach has to determine costs for faulty decisions (and Neyman-Pearson do this implicitly, by choosing the level α of a test). This involves, however, reference to the *purposes* to which we want to put our newly acquired knowledge. For Fisher, this is not compatible with the idea of science as pursuit of truth. Statistical inference has to be "convincing to all freely reasoning minds, entirely independently of any intentions that might be furthered by utilizing the knowledge inferred" (Fisher 1956, 103). Second, in science, a judgment on the truth of a hypothesis is usually not made

on the basis of a single experiment. Instead, we obtain some *provisional* result which is refined through further analysis. By their behavioral rationale and by making a "decision" between two hypotheses, Neyman and Pearson insinuate that the actual data justify a judgment on whether H_0 or H_1 is true. Such judgments have, according to Fisher, to be suspended until further experiments confirm the hypothesis, ideally using varying auxiliary assumptions and experimental designs. Third, Neyman and Pearson test a statistical hypothesis against a definite alternative. This leads to some results that appear paradoxical. Take, for instance, the example of a normal distribution with known variance $\sigma^2 = 1$ where the hypothesis about the mean $H_0 : \mu = 0$ is tested against the hypothesis $H_1 : \mu = 1$. If the average of the observations centers, say, around -5, it appears that neither H_0 or H_1 should be 'accepted'. Nevertheless, the Neyman-Pearson rationale contends that in such a situation we have to accept H_0 because the discrepancy to the actual data is less striking than with H_1. In such a situation, when H_0 offers a poor fit to the data, such a decision is arguably weird.

Summing up, Fisher disqualifies Neyman and Pearson's decision-theoretic approach as a mathematical "reinterpretation" of his own significant tests that is utterly inappropriate for use in the sciences – he even suspects that Neyman and Pearson would not have come up with their approach had they had "any real familiarity with work in the natural sciences" (Fisher 1956, 76). Therefore he developed a methodology of his own which proved to be extremely influential in the natural as well as in the social sciences. His first two books, "Statistical Methods for Research Workers" (1925) and "The Design of Experiments" (1935) quickly went through many reprints and shaped the applications of statistics in the sciences for decades. The core of his method is the *test of a point null hypothesis* or *significance test*. Here, we want to tell chance effects from real effects. To this end, we check whether a null (default, chance) hypothesis is good enough to fit the data. For instance, we want to test the effects of a new learning device on students' performance, and we start with the default assumption that the new device yields no improvement. If that hypothesis is apparently incompatible with the data (if the results are 'significant'), we conclude that there is some effect in the treatment. The core of the argument consists in *'Fisher's Disjunction'*:

> "Either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true."(Fisher 1956, 39)

In other words, the occurrence of a result that is very unlikely to be a product of mere chance (students using the device scoring much better than the rest) strongly speaks against the null hypothesis that there is no effect. Significant findings under the null suggest that there is more than pure chance involved, that there is some kind of systematic effect going on. As we will see below, this disjunction should be regarded with great caution, and it has been the source of many confusions and misunderstandings.

Figure 1 illustrates the difference between Neyman-Pearsonian and Fisherian tests for the case of testing hypotheses on the mean value of a Normal distribution. The probability

$$p := P_{H_0}(T(X) > T(x_0)) \tag{1}$$

gives the *level of significance* which the observed value x_0 achieves under H_0, with respect to a function T that measures distance from the null hypothesis

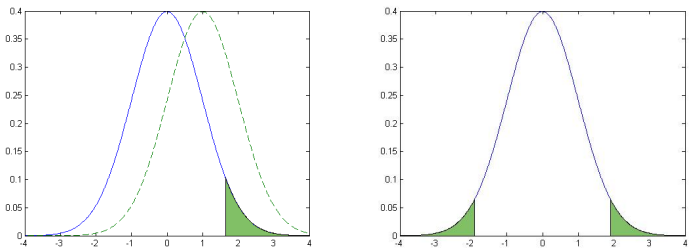

Figure 1: Left figure: The null hypothesis $H_0 : N(0,1)$ (full line) is tested at the 5%-level against against the alternative $H_0 : N(1,1)$ (dashed line). Right figure: a Fisherian significance test of H_0 against an unspecified alternative. – The shaded areas represents the set of results where H_0 is rejected in favor of H_1, respectively where the results speak "significantly" against H_0.

H_0. p is also often called the *p-value* induced by x_0, and is supposed to give a rough idea of the tenability of the null. The higher the discrepancy, the more significant the results.

The rationale underlying Fisher's Disjunction displays striking similarity to Karl Popper's falsificationist philosophy of science: A hypothesis H_0, which should be as precise and free of ambiguity as possible, is tested by checking its observational implications. If our observations contradict H_0, we reject it and replace it by another hypothesis. However, this understanding of falsificationism only applies to testing deterministic hypotheses. Observations are never incompatible with probabilistic hypotheses, they are just very unlikely. Therefore Popper (1959, 191) expanded the falsificationist rationale by saying that we regard a hypothesis H_0 as false when the observed results are improbable enough. This is exactly the rationale of Fisher's Disjunction. Notably, Fisher formulated these ideas as early as Popper and independently of him. The methodological similarity between Popper and Fisher's views becomes even more evident in the following quote:

> "[...] 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 1935, 19)

This denial of positive confirmation of the null by non-significant results fits not only well with Popper's view on confirmation and corroboration, but also with a more modern textbook citation:

> "Although a significant departure [from the null] provides some degree of evidence against a null hypothesis, it is important to realize that a 'nonsignificant' departure does not provide positive evidence in favor of that hypothesis. The situation is rather that we have failed to find strong evidence against the null hypothesis." (Armitage and Berry 1987, 96)

Thus, the symmetry of the Neyman-Pearsonian approach is broken: While Neyman-Pearson tests end up "accepting" either hypothesis (and building action

on the basis of this decision), Fisherian significance tests understand a significant result as strong evidence against the null hypothesis, but an insignificant result does not mean evidence for the null.

The attentive reader might have noticed that Fisher's Disjunction is actually inconsistent with his own criticism of the Neyman-Pearson appraoch. Recall that Fisher argued that significant outcomes do not deliver final verdicts on the feasibility of the null hypothesis. Rather, they state *provisional* evidence against the null. But how is this compatible with the idea of "disproving the null" by means of significance tests? To reconcile both positions, Fisher has to admit some abuse of language:

> "[...] if we use the term rejection for our attitude to such a [null] hypothesis, it should be clearly understood that no irreversible decision has been taken; that as rational beings, we are prepared to be convinced by future evidence that [...] in fact a very remarkable and exceptional coincidence had taken place." (Fisher 1959, 35)

In the light of these ambiguities, it does not surprise that Fisher's writings have been the source of many misunderstandings, and that scientists sometimes use fallacious practices or interpretations while believing that these practices have been authorized by a great statistician. Before describing the problems of significance tests, however, I would like to shed light on the contrast between frequentist statistics, which comprises Fisher's approach as well as the Neyman-Pearson paradigm, and the rivalling school of Bayesian statistics.

6 Fisher versus Bayes

Bayesian inference is a school of statistics with great significance for some theoretical branches of the social sciences, such as decision theory, game theory and the psychology of human reasoning. The principles of Bayesian inference are explained in the chapter on decision theory, so we restrict ourselves to a brief outline of the basic idea. Suffice to say that Bayesian statistics is, essentially, a theory of *belief revision*: prior beliefs on the credibility of a hypothesis H are represented by mathematical probabilities, modified in the light of incoming evidence E and transformed into posterior beliefs (represented by a conditional probability, $P(H|E)$). The relevant formula that expresses how these beliefs are changed is Bayes's Theorem:

$$\begin{aligned} P(H|E) &= \frac{P(H)\,P(E|H)}{P(E)} = \frac{P(H)\,P(E|H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)} \\ &= \left(1 + \frac{1-P(H)}{P(H)}\,\frac{P(E|\neg H)}{P(E|H)}\right)^{-1}. \end{aligned}$$

Thus, the sampling distributions of E under H and $\neg H$ are combined with the prior probability of H in order to arrive at a comprehensive verdict on the credibility of H in the light of evidence E.

Modern philosophers of statistics – but also scientists themselves – have stressed the contrast between frequentist and Bayesian inference, depicting them as mutually exclusive schools of statistics (Howson and Urbach 2006; Mayo 1996). The polemics which both Bayesians and frequentists use to mock at

their respective opponents adds to the image of statistics as a deeply divided discipline where two enemy camps are quarreling about the right foundations of inductive inference. In particular, Bayesians have been eager to point out the limitations and shortcomings of frequentist inference for scientific applications, such as in the seminal paper of Edwards, Lindman and Savage (1963). Notably, this influential methodological contribution appeared not in a statistics journal, but in *Psychological Review*! On the other hand, frequentist criticisms of Bayesian inference read equally harshly.

These heated debates do not do justice to the intentions of the founding fathers, who are often more pragmatic than one might be inclined to think in retrospect. Take the case of Ronald A. Fisher. Although Fisher is correctly perceived as one of the founding fathers of frequentist inference, it is wrong to see him as an anti-Bayesian. True, Fisher objects to the use of prior probabilities in scientific inference. But it is important to see why and under which circumstances. In principle, he says, there is nothing wrong with using Bayes' formula to revise one's belief, it is just *practically* impossible to base a sound scientific judgment on them. For how shall we defend a specific assignment of prior beliefs vis-à-vis our fellow scientists if they are nothing more than psychological tendencies? Most often, there is no knowledge available on which we could base specific prior beliefs (1935, 6-7; 1956, 17). That said, Fisher speaks very respectfully about Bayes and his framework: Bayesian inference may be appropriate in science if genuine prior knowledge is available (1935, 13), and he admits the rationality of the subjective probability interpretation in spite of his own inclination to view probabilities as relative frequencies.

It is therefore important to note that the debate between frequentist (here: Fisherian) and Bayesian statistics is not in the first place a debate about the principles of inductive inference *in general*, but a debate about which kind of inference is more appropriate for the purposes of *science*. The following section will cast some doubts on the appropriateness of pure, unaided significance testing in the social sciences.

7 The Pitfalls of Significance Testing

The practice of significance tests has been dominating experiments in the social sciences for more than half a century. Journal editors and referees ask for significance tests and p-values (quantities describing the level of significance), standardizing experimental reports in a wide variety of branches of science (econometrics, experimental psychology, behavioral economics, etc.). Alternative approaches, e.g. the application of Bayesian or likelihoodist statistics to the evaluation of an experiment, had few chances of being published.

These publication practices in the last decades are at odds with the existence of a long methodological debate on significance testing in the social sciences (e.g. Rozeboom 1960). In that debate, statisticians and social scientists – mostly mathematically educated psychologists – have repeatedly criticized the misuse of significance tests in evaluating and interpreting scientific experiments. Before going into the details of that debate, we briefly list some apparent advantages of significance testing.

Objectivity Significance tests avoid the subjective probabilities of Bayesian statistics. Thereby the observed levels of significance seem to be an objec-

tive standard for evaluating the experiment, e.g. for telling a chance effect from a real one.

No Alternative Hypotheses Significance tests are a means of testing a single, exact hypothesis, without specifying a certain direction of departure (i.e. an alternative hypothesis). Therefore, significance tests detect more kinds of deviation from that hypothesis than Neyman-Pearson tests do.

Replicability Significance tests address the issue of replicability – namely the significance level can be understood as the relative frequency of observing a more extreme result if (i) the null hypothesis were true and (ii) the trial were repeated very often.

Practicality Significance tests are easy to implement, and significance levels are easy to compute.

However, it is not clear whether these advantages of significance tests are really convincing. We discuss a couple of objections.

Fisher's Disjunction revisited. The original example which Fisher used to motivate his famous disjunction was the hypothesis that the stars are evenly distributed in the sky, i.e. the chance that a star is in a particular area of the sky is proportional to the size of that area. Thus, if there are a lot of stars next to a particular star, such an event is unlikely to happen due to chance. Indeed, clusters of stars are frequently observed. We may, according to Fisher's Disjunction, rule out the hypothesis of uniform distributions and conclude that stars tend to cluster.

However, Hacking (1965, 81-82) has convincingly argued that such an application of Fisher's Disjunction is fallacious. Under the hypothesis of uniform distribution, every constellation of stars is extremely unlikely, and there are no likely vs. unlikely chances, but *only* 'exceptionally rare chances'. If Fisher's Disjunction were correct, we would thus always have to reject the hypothesis of uniform distribution, independent of the outcome. This amounts to a *reductio* of significance testing since clearly, hypotheses that postulate a uniform distribution are testable, and they often occur in scientific practice.

To circumvent Hacking's objection, we might interpret Fisher's Disjunction in a different way. For instance, we could read the 'exceptionally rare chance' as a chance that is exceptionally rare *compared to other possible events*, instead of 'a probability lower than a fixed value p'. Still, this does not help us in the present problem because the uniform distribution postulates that all star constellations are equally likely or unlikely. Thus, the notion of a relatively rare chance ceases to apply (Royall 1997, 65-68).

One might concede Hacking's objection for this special case and try to rescue significance tests in general by introducing a parameter of interest ϑ. This is a standard situation in statistical practice. For instance, let's take a coin flip model which has "heads" and "tails" as possible outcomes, and ϑ denotes the propensity of the coin to come up heads. Under the null hypothesis $H_0 : \vartheta = 0.5$, all sequences of heads and tails are equally likely, but still, it is ostensibly meaningful to say that 'HHHHHTTTTT' or 'THTHTHTHTH' provides less evidence against H_0 than 'HHHHHHHHHH' does. The technical concept for implementing this intuition consists in calculating the chance of a transformation of the data that is a *minimally sufficient* statistic with respect to the parameter of

interest ϑ, such as the number of heads or tails. Then we get the desired result that ten heads, but not five heads vs. five tails (in whatever order) constitute a significant finding against H_0. Thus, there is no exceptionally rare chance as such – any such chance is relative to the choice of a parameter that determines *the way in which the data are exceptional.*

This line of reasoning fits well with the above example, but it introduces implicit alternative hypotheses. When relativizing unexpectedness to a parameter of interest, we are committing ourselves to a specific class of potential alternative hypotheses – namely those hypotheses that correspond to the other parameter values. When applying Fisher's Disjunction, we do not judge the tenability of H_0 'in general', without recourse to a specific parameter of interest or a class of alternatives – we always examine a certain way the data could deviate from the null. Thus, we are not testing the probability model H_0 as such, but a particular aspect thereof, such as 'why that value of ϑ rather than another one?'. The choice of a parameter reveals a class of intended alternatives.[12]

This has some general morals: what makes an observation evidence against a hypothesis is not its low probability under this hypothesis, but its low probability compared to an alternative hypothesis. An improbable event is not evidence against a hypothesis per se, but

> "[...] what it does show is that if there is any alternative hypothesis which will explain the occurrence of the sample with a more reasonable probability [...] you will be very much more inclined to consider that the original hypothesis is not true." (William S. Gosset ("Student") in private communication to Egon Pearson, quoted in Royall 1997, 68.)

Thus, Fisher's Disjunction and the inference from relatively unlikely results to substantial evidence is caught in a dilemma: Either we run into the inconsistencies described above, or the choice of the test statistic reveals implicit alternatives to which the hypothesis is compared. Then the falsificationist heuristics of Fisher's Disjunction has to be replaced by an account of contrastive testing. Then, it is unclear to what extent the Fisherian framework of significance testing can claim any advantage vis-à-vis Neyman and Pearson's tests of two competing hypotheses.

The Base Rate Fallacy. Gigerenzer (1993) famously characterized the inner life of a scientist who uses statistical methods by means of an analogy from psychoanalysis: there is a Neyman-Pearsonian Super-Ego, a Fisherian Ego and a Bayesian Id. The Neyman-Pearsonian Super-Ego preserves a couple of unintuitive insights, e.g. that we cannot test a theory without specifying alternatives, that significance tests only give us the probability of data given a hypothesis, instead of an assessment of the hypothesis's credibility. The Bayesian Id is located at the other end of the spectrum, incorporating the researcher's desire for posterior probabilities of a hypothesis, as a measure of its tenability or credibility. The Ego is caught in the conflict between these extremes, and acts as the scientist's guide through reality. It adopts a Fisherian position where both

[12] There is no canonical class of alternatives: we could plausibly suspect that the coin has an in-built mechanism that makes it come up with alternating results, and then, 'THTHTHTHTH' would not be an insignificant finding, but speak to a high degree against the chance hypothesis.

extremes are kept in balance: significance test neither give behavioral prescriptions nor posterior probabilities, rather they yield "a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test" (Fisher 1956, 44).

However, the Bayesian Id sometimes breaks through. As pointed out by Oakes (1986) and Gigerenzer (1993), most active researchers in the social sciences – even those with statistical education – tend to interpret significance levels (e.g. $p = 0.01$) as posterior probabilities of the null hypothesis, or at least as overwhelming evidence against the null. Why is this inference wrong?

Assume that we want to test a certain null hypothesis against a very implausible alternative, e.g. that the person under test has a very rare disease. So the null denotes absence of that disease. Now, a highly sensitive test, that is right about 99.9% of the time, indicates presence of the disease, yielding a very low p-value. Many people would now we tempted to conclude that the person probably has the disease. But since that disease is rare, the posterior probability of the null hypothesis can still be very large. In other words, *evidence* that speaks to a large degree against the null is not sufficient to support a *judgment* against the null – it would only do so if the null and the alternative were about equally likely at the outset. Such a failure to recognize the dependency between the *base rate* of the null hypothesis and the strength of the final evidential judgment is called the *base rate fallacy.*

Although that fallacy is severe and widespread (and similar misinterpretations of significance tests abound, see Gigerenzer 2008), those fallacies might speak more against the *practice* of significance testing than against significance tests themselves. In any case, they invite to misinterpretations, especially because p-values (significance levels) are hard to related to scientifically meaningful conclusions.[13]

The Replicability Fallacy. This fallacy is more subtle than the base rate fallacy. It does not interpret p-values as posterior probabilities, but understands a p-value of, say, 0.05 as saying that if the experiment were repeated, a result that was at least as significant as the present observations would occur at 95% of the time. Thus, the outcome is believed to have implications for the recurrence of a significant result and for the replicability of the present observations. And replicability is, needless to say, one of the main quality brands of good experiments.

In principle, there is nothing wrong with connecting replicability to significance testing. But a crucial premise is left out – namely that the replication frequency holds only *under the assumption that the null hypothesis is true.* Since the power of many significance tests is low, implying that nonsignificant results often occur when the null is actually false, the kind of replicability that significance tests ensure is much more narrow than desired (Schmidt and Hunter 1997). A solution to this problem that has gained more and more followers in the last decades is to replace significance levels by confidence intervals that address the issue of replicability regardless of whether the null hypothesis is actually true.

The Jeffrey-Lindley Paradox. This problem sheds light on the importance of sample size in statistical testing, and applies to both Fisher's and

[13]See Casella and Berger (1987) and Sellke and Berger (1987) for more detailed discussions of the evidential value of p-values in different testing problems.

Neyman and Pearson's framework. For a large enough sample, a point null hypothesis can be rejected at a significant level while the posterior probability of the null approaches one (Lindley 1957). Take, for instance, a normal model $N(0,1)$ where we test the value of the mean: $H_0 : \mu = 0$ against an alternative $H_1 : \mu = 1$. Since the sampling distribution of the mean of n samples $\overline{X_n}$ approaches $N(0, 1/n)$, any slight deviation of the mean from the null hypothesis will suffice to make the result statistically significant. Even more, if we decide to sample on until we get significant results against the null hypothesis, we will finally get them (Mayo and Kruse 2001).

At the same time, the posterior of the null hypothesis also converges to 1 with increasing n, as long as the divergence remains rather small. Thus, for large samples, significance levels do not reliably indicate whether or not a certain effect is present, and can grossly deviate from the hypothesis's posterior credibility. Significance tests may tell us whether there is evidence against a *point* null hypothesis, but they don't tell us whether that effect is large enough to be of scientific interest.

Statistical versus practical significance. Typically, the null hypothesis typically denotes an idealized hypothesis, such as "there is no difference between the effects of A and B". In practice, no one believes such a hypothesis to be literally true, rather, everyone expects that there are differences, but perhaps just at a minute degree: "The effects of A and B are always different – in some decimal place – for some A and B. Thus asking 'Are the effects different?' is foolish." (Tukey 1991, 100)

However, even experienced scientists often read tables in an article by looking out for asterisks: one asterisk denotes "significant" findings ($p < 0.05$), two asterisks denote "highly significant" ($p < 0.01$) findings. It is almost impossible to resist the psychological drive to forget about the subtle differences between statistical and scientific significance, and many writers exploit that fact:

> "All psychologists know that statistically significant does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results sections studded with asterisks becomes in the Discussion section highly significant or very highly significant, important, big!" (Cohen 1994, 1001)

Instead, statistical significance should at best mean that evidence speaks against our idealized hypothesis while we are still unable to give the direction of departure or the size of the observed effect (Kirk 1996). This provisional interpretation is in line with Fisher's own scepticism regarding the interpretation of significance tests, and Keuzenkamp and Magnus's (1995) observation that significance testing in econometrics rarely leads to the dismissal of an economic theory, and its subsequent replacement.

Finally, under the assumption that null hypotheses are strictly spoken wrong, it is noteworthy that significance tests bound the probability of erroneously rejecting the null, while putting no constraints on the probability of erroneously *accepting* the null, i.e. the power of a test. Considerations of power, sample size and effect size that are fundamental in Neyman and Pearson's approach fall out of the simplified Fisherian picture of significance testing. This is not to say that these tests are worthless – for instance, in econometrics, a series of significance tests can be very useful to detect whether a model of a certain process has been *misspecified*. Significance tests look for directions in different departures

(autocorrelation, moving average, etc.), and significant results provide us with reasons to believe that our model has been misspecified, and make us think harder about the right form of the model that we want to use in future research (Mayo and Spanos 2004; Spanos 1986).

In that spirit, it should be stressed once more that Fisher considered significance tests to be a preliminary, explorative form of statistical analysis that gives rise to further investigations, not to final decisions on a hypothesis. But reading social science journals, it is not always clear that the practicing researchers are aware of the problem. The penultimate section briefly sketches how this problem was addressed in the last decades.

8 Recent Trends

All the criticisms of significance testing have led many authors to conclude that significance tests do not help to address scientifically relevant questions. Using them in spite of their inability to address the relevant questions only invites to misuse and confusion (Cohen 1994; Schmidt 1996). Since the problem and its discussion was especially pronounced in experimental psychology, we focus on the reactions in that field.

Recognizing that those criticisms were justified, the American Psychological Association (APA) appointed a Task Force on Statistical Interference (TFSI) whose task consisted in investigating controversial methodological issues in inferential statistics, including significance testing and its alternatives (Harlow et al. 1997; Thompson 1999a; Wilkinson et al. 1999). After long deliberation, the Task Force came up with some recommendations that made the APA change their publication guidelines, affecting major journals affiliated to the APA, such as *Psychological Review*. The commission stated, for instance, that p-values do not reflect the significance or magnitude of an observed effect, and "encouraged" authors to provide information on effect size, either by means of directly reporting an effect size measure (e.g. Pearson's correlation coefficient r or Cohen's effect size measure d), or power and sample size of the test.

However, as predicted by Sedlmeier and Gigerenzer (1989) and observed by a lot of empirical studies on research practice (e.g. Keselman et al. 1998), the admonitions and encouragements of the APA publication manual proved to be futile. First, psychologists were not trained at computing and working with effect sizes. Second, "there is only one force that can effect a change, namely the editors of the major journals" (Sedlmeier and Gigerenzer 1989, 315). Encouragement was likely to be ignored when compared to the *compulsory* requirements when submitting a manuscript and abiding by formatting guidelines:

> "To present an 'encouragement' in the context of strict absolute [manuscript] standards [...] is to send the message 'these myriad requirements count, this encouragement doesn't'." (Thompson 1999b, 162)

However, the extensive methodological debate finally seems to bear fruit. As pointed out by Vacha-Haase et al. (2000), quite some editors changed their policy, requiring the inclusion of effect size measures, where unwillingness to comply with that guideline had to be justified in a special note. This development, though far from overturning and eliminating all fallacious practices,

shows that sensitivity for the issue has increased, and raises hope for the future.

Also, Bayesian methods (and other approaches, such as Royall's (1997) likelihoodism) gain increasing acceptance beyond purely technical journals. Such inferential methods can now, to an increasing extent, be found in major psychology journals as well. Finally, there is an increasing amount of journals that address a readership that is interested in mathematical and statistical modeling in the social sciences, as well as in methodological foundations. Although the presentation and interpretation of statistical findings in the social sciences is still wanting, there is some reason for optimism: the problems have been discovered and addressed, and we are now in the phase where a change towards a more reliable methodology is about to be effectuated. As stated by Cohen (1994), this change is slowed down by the conservativeness of many scientists, and their desire for automated inferential mechanisms. But such "cooking recipes" do, as the drawbacks of significance tests teach us, not exist.

9 Summary

Let us conclude. In this contribution, we have surveyed and classified a variety of mathematical methods that are used in the social sciences and argued that such techniques, in spite of several methodological objections, can add extra value to social scientific research. Then, we have focused on methodological issues in statistics – the part of mathematics that is most frequently used in the social sciences, in particular in the design and interpretation of experiments. We have represented the emergence of and the rationale behind the ubiquitous significance tests, as well as explained the pitfalls to which many researches fall prey when using them. Finally, after comparing significance testing to rivalling schools of statistical inference, we have discussed recent trends in the methodology of the social sciences and argued that there is reason for optimism, and that awareness of methodological problems, as well as interest for mathematical and statistical techniques is growing.

References

Anderson, P., K. Arrow and D. Pines (eds.) (1988): *The Economy As An Evolving Complex System.* Redwood City: Addison-Wesley.

Armitage, P., and G. Berry (1987): *Statistical Methods in Medical Research.* Second Edition. New York: Springer.

Arrow, K. et al (eds.) (1960): *Mathematical Methods in the Social Sciences.* Stanford: Stanford University Press.

Backhouse, R. (1995): *A History of Modern Economic Analysis.* Oxford: Blackwell.

Balzer, W., C.U. Moulines and J. Sneed (1987): *An Architectonic for Science: The Structuralist Program.* Dordrecht: Reidel.

Balzer, W. and B. Hamminga (eds.) (1989): *Philosophy of Economics*. Dordrecht: Kluwer.

Beed, C. and O. Kane (1991): What Is the Critique of the Mathematization of Economics? *Kyklos* 44 (4), 581–612.

Beisbart, C. and S. Hartmann (2009): Welfarist Evaluations of Decision Rules under Interstate Utility Dependencies, *Social Choice and Welfare* doi: 10.1007/s00355-009-0399-z.

Berger, J.O., and T. Sellke (1987). Testing a point null hyphoteses: The irreconciliability of p-values and evidence (with discussion). *Journal of the American Statistical Association* 82, 112–122.

Bermúdez, J.L. (2009): *Decision Theory and Rationality*. Oxford: Oxford University Press.

Casella, G., and R. L. Berger (1987): Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association* 82, 106-111.

Cohen, J. (1994): The Earth is round ($p < .05$). *American Psychologist* 49, 997–1001.

Debreu, G. (1991): The Mathematization of Economic Theory. *American Economic Review* 81(1), 1–7.

Edwards, W., H. Lindman and L.J. Savage (1963): Bayesian Statistical Inference for Psychological Research. *Psychological Review*, 70, 450-499.

Fisher, R.A. (1925): *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1935): *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Fogel, R.W. (1975): The Limits of Quantitative Methods in History, *American Historical Review* 80(2): 329–50.

Frigg, R. and S. Hartmann: Models in Science, in: *The Stanford Encyclopedia of Philosophy*, (Spring 2006 Edition).

Frigg, R. and J. Reiss (2009). The Philosophy of Simulation: Hot New Issues or Same Old Stew? *Synthese* 169 (3).

Gaertner, W. (2006): *A Primer in Social Choice Theory*. New York: Oxford University Press.

Gallegatti, M., S. Keen, T. Lux, and P. Ormerod (2006): Worrying Trends in Econophysics, Physica A. *Physica A* 370(1), 1-6.

Gilbert, N. and K. Troitzsch (2005): *Simulation for the Social Scientist*. New York: McGraw-Hill.

Gigerenzer, G. (1987): Probabilistic Thinking and the Fight against Subjectivity, in Krüger et al. (eds., 1987): The Probabilistic Revolution, 11–33.

Gigerenzer, G. (1993): The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*, 311–339. Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (2008): *Rationality for Mortals: How Humans Cope with Uncertainty.* Oxford: Oxford University Press.

Goodman, S. (1999): Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine* 130, 995-1004.

Granger C. (1999): *Empirical Modelling in Economics: Specification and Evaluation.* Cambridge: Cambridge University Press.

Grodon, S. (1991): *The History and Philosophy of Social Science.* London: Routledge.

Hacking, I. (1965): *Logic of Statistical Inference.* Cambridge: Cambridge University Press.

Harlow, L.L., S.A. Mulaik, and J.H. Steiger (eds.) (1997): *What if there were no significance tests?* Mahwah/NJ: Erlbaum.

Hands, W. (2008): Popper and Lakatos in Economic Methodology. In: D. Hausman (ed.), *The Philosophy of Economics: An Anthology.* Cambridge: Cambridge University Press, 188–203.

Hartmann, S. (1996): The World as a Process: Simulations in the Natural and Social Sciences, in: R. Hegselmann et al. (eds.), *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View.* Dordrecht: Kluwer, 77–100.

Hartmann, S. and W. Meijs (2010): Walter the Banker: The Conjunction Fallacy Reconsidered, to appear in *Synthese.*

Hausman, D. (1992): *The Inexact and Separate Science of Economics.* Cambridge: Cambridge University Press.

Hausman, D. (ed.) (2008): *The Philosophy of Economics: An Anthology.* Cambridge: Cambridge University Press.

Hegselmann et al. (eds.) (1996): *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View.* Dordrecht: Kluwer.

Hertwig, R., B. Benz and S. Krauss (2008): The Conjunction Fallacy and the many Meanings of *and. Cognition*, 108: 740–753.

Howson, C., and P. Urbach (2006): *Scientific Reasoning: The Bayesian Approach.* Third Edition. Open Court, La Salle.

Humphreys, P. (2004): *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford: Oxford University Press.

Jackson, M. (2008): *Social and Economic Networks.* Princeton: Princeton University Press.

Keselman, H.J., et al. (1998): Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research* 68, 350–386.

Keuzenkamp, H., and J. Magnus (1995): On tests and significance in econometrics. *Journal of Econometrics* 67, 5–24.

Kirk, R. (1996): Practical Significance: A concept whose time has come. *Educational and Psychological Measurement* 56: 746–759.

Kolmogorov, A. N. (1933/56): *Foundations of the Theory of Probability.* Original work published in German in 1933. New York: Chelsea.

Krantz, D.H., R.D. Luce, P. Suppes, and A. Tversky (1971): *Foundations of Measurement. Vol. I. Additive and Polynomial Representations.* New York: Academic Press.

Krüger, L., G. Gigerenzer, and M. Morgan (eds.) (1987): *The Probabilistic Revolution, Vol. 2: Ideas in the Sciences.* Cambridge/MA: The MIT Press.

Lindley, D. (1957): A Statistical Paradox. *Biometrika* 44, 187–192.

List, C. and C. Puppe (2009): Judgment Aggregation: A Survey. In: P. Anand, C. Puppe and P. Pattaniak (eds.), *Oxford Handbook of Rational and Social Choice.* Oxford: Oxford University Press, 457–482.

Luce, R.D. and P. Suppes (1968): Mathematics. In: *International Encyclopedia of the Social Sciences*, Vol. 10. New York: Macmillan and Free Press, 65–76.

Luce, R.D., D.H. Krantz, P. Suppes, and A. Tversky (1990). *Foundations of measurement. Vol. III. Representation, Axiomatization, and Invariance.* New York: Academic Press.

Mantegna, R. and H. Stanley (1999): *An Introduction to Econophysics: Correlations and Complexity in Finance.* Cambridge: Cambridge University.

Marchi, S. de (2005): *Computational and Mathematical Modeling in the Social Sciences.* Cambridge: Cambridge University Press.

Mayo, D.G. (1996): *Error and the Growth of Experimental Knowledge.* Chicago: Chicago University Press.

Mayo, D.G., and M. Kruse (2001): Principles of inference and their consequences, in: D. Cornfield, J. Williamson (eds.): *Foundations of Bayesianism*, 381-403. Dordrecht: Kluwer.

Mayo, D.G., and A. Spanos (2004): Methodology in Practice: Statistical Misspecification Testing. *Philosophy of Science* 71: 1007-1025.

McCauley, J. (2004): *Dynamics of Markets: Econophysics and Finance.* Cambridge: Cambridge University Press.

Mirowski, P. (1990): *More Heat than Light: Economics as Social Physics, Physics as NatureÕs Economics.* Cambridge: Cambridge University Press.

Mirowski, P. (2001): *Machine Dreams: Economics Becomes a Cyborg Science.* Cambridge: Cambridge University Press.

Morgan, M. (1987): Statistics without Probability and Haavelmo's Revolution in Econometrics, in Krüger et al. (eds., 1987): The Probabilistic Revolution, 171–197.

Morgan, M. (2002): *The History of Econometric Ideas.* Cambridge: Cambridge University Press.

Morrow, J. (1994): *Game Theory for Political Scientists.* Princeton: Princeton University Press.

Neyman, J., and E. Pearson (1933): On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A* 231, 289–337.

Neyman, J., and E. Pearson (1967): Joint Statistical Papers. Cambridge: Cambridge University Press.

Oakes, M. (1986): *Statistical Inference. A commentary for the social and behavioral sciences.* New York: Wiley.

Piotrowski, E. and J. Sladkowski (2003): An Invitation to Quantum Game Theory, *International Journal of Theoretical Physics* 42 (5): 1089–1099.

Popper, K. R. (1959): *The Logic of Scientific Discovery.* London: Routledge.

Rosenberg, A. (1976): *Microeconomic Laws: A Philosophical Analysis.* Pittsburgh: University of Pittsburgh Press.

Rosenberg, A. (1992): *Economics – Mathematical Politics or Science of Diminishing Returns.* Chicago: University of Chicago Press.

Royall, R. (1997): *Statistical Evidence – A Likelihood Paradigm.* London: Chapman & Hall.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test, *Psychological Bulletin* 57, 416–428.

Schmidt, F.L. (1996): Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* 1, 115–129.

Schmidt, F.L., and J.E. Hunter (1997): Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data, in Harlow et al. (eds.), What if there were no significance tests?, 37-64.

Schultz, H. (1928): *Statistical Laws of Demand and Supply with Special Application to Sugar.* Chicago: University of Chicago Press, Chicago.

Sedlmeier. P., and G. Gigerenzer (1989): Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105, 309–316.

Sen, A (1999): The Possibility of Social Choice, *The American Economic Review* 89(3), 349–378.

Senn, P. (1971): *Social Science and Its Methods.* Boston: Holbrook Press.

Simon, H. (1996): *Models of my Life.* Cambridge: MIT Press.

Spanos, A. (1986). *Statistical foundations of econometric modelling.* Cambridge: Cambridge University Press.

Sugden, R. (2008): Credible Worlds: The Status of Theoretical Models in Economics. In: D. Hausman (ed.), *The Philosophy of Economics: An Anthology.* Cambridge: Cambridge University Press, 476–509.

Suppes, P. (1967): What is a Scientific Theory? In: S. Morgenbesser (ed.): *Philosophy of Science Today.* New York: Basic Book, 55–67.

Suppes, P., D.H. Krantz, R.D. Luce, and A. Tversky (1989): *Foundations of Measurement. Vol. II. Geometrical, Threshold and Probabilistic Representations.* New York: Academic Press.

Suppes, P. (2001): Representation and Invariance of Scientific Structures. Chicago: University of Chicago Press.

Thompson, B. (1999a): If statistical significance tests are broken/misused, what practice should supplement or replace them? *Theory & Psychology* 9: 167–183.

Thompson, B. (1999b): Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review* 11: 157–169.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science* 6, 100-116.

Tversky, A. and D. Kahneman (1983): Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review* 90: 293–315.

Vacha-Haase, T., and al. (2000): Reporting Practices and APA Editorial Policies Regarding Statistical Significance and Effect Size. *Theory & Psychology* 10, 413–425.

Vazquez, A. (1995): Marshall and the Mathematization of Economics. *Journal of the History of Economic Thought* 17, 247–265.

Voit, J. (2000): *The Statistical Mechanics of Financial Markets.* Berlin: Springer.

Waldrop, M. (1992): *Complexity: The Emerging Science at the Edge of Order and Chaos.* New York: Simon & Schuster.

Wallace, D. (2010): A Formal Proof of the Born Rule from Decision-Theoretic Assumptions. To appear in: S. Saunders et al (eds.), *Many Worlds? Everett, Quantum Theory, and Reality.* Oxford: Oxford University Press.

Weidlich, W. (2006): *Sociodynamics: A Systemic Approach to Mathematical Modelling in the Social Sciences.* New York: Dover Publications.

Weintraub, R. (2002) : *How Economics Became a Mathematical Science.* Durham: Duke University Press.

Wigner, E.P. (1967): *Symmetries and Reflections.* Bloomington: University of Indiana Press.

Wilkinson, L., and Task Force on Statistical Inference (1999): Statistical Methods in Psychology Journals: Guidelines and Explanations, *American Psychologist.* Vol. 54, Aug 1999, 594–604.