Revista Sergipana de Matemática e Educação Matemática

https://doi.org/10.34179/revisem.v6i3.14904

THE MEANING OF PROBABILITY FROM A FOUNDATIONAL PERSPECTIVE

Manfred Borovcnik
University of Klagenfurt, Klagenfurt, Austria.
manfred.borovcnik@aau.at

Abstract

We focus on the ways in which we can use a frequentist interpretation of probability to develop suitable methods for statistical inference. The discussion about the controversy in the foundations reveals that a frequentist conception is highly prone to dispute, as a justification of this view fails from a rational perspective when the explication of probability integrates statistical inference. We give an overview on the dispute and the crucial examples that highlight the deficiencies of a purely frequentist position towards probability. The concept of probability emerges from a mixture of classical, frequentist, and subjectivist meanings, which are not easy to separate. A shift in connotation of probability towards a biased frequentist meaning decreases the scope of probability or the quality of applications. Probability is a complementary concept, which falls apart if we reduce it to one view. This gives rise to investigate refined approaches towards teaching from a wider perspective on the range of meanings of probability apart from frequentist aspects. Empirical studies show the shortcomings of educational approaches that ignore subjectivist aspects of probability, which leads to far-reaching misconceptions not only about the use of Bayes' formula but also in the perception of probabilities at large.

Resumo

Nesse texto temos como foco a discussão sobre as maneiras pelas quais uma interpretação frequentista da probabilidade pode ser usada para desenvolver métodos adequados para inferência estatística. O debate sobre a controvérsia sobre os fundamentos revela que uma concepção frequentista é altamente propensa a argumentos como uma justificativa desse ponto de vista falha em uma perspectiva racional quando a explicação da probabilidade integra a inferência estatística. Damos uma visão geral do debate e dos exemplos cruciais que destacam as fragilidades de uma posição puramente frequentista em relação à probabilidade. O conceito de probabilidade emerge de uma mistura de significados clássicos, frequentistas e subjetivistas, que não são fáceis de separar. A mudança na conotação da probabilidade para um significado frequentista tendencioso

diminui seu escopo ou a qualidade das aplicações. A probabilidade é um conceito complementar que se perde força se o reduzirmos a uma única concepção. Isto dá origem a pesquisas sobre abordagens para o ensino de uma perspectiva mais ampla sobre a gama de significados de probabilidade além dos aspectos frequentistas. Estudos empíricos mostram as deficiências das abordagens educacionais que ignoram os aspectos subjetivistas da probabilidade, o que leva a graves equívocos não só sobre o uso da fórmula de Bayes, mas também sobre a percepção das probabilidades em geral.

1 Introduction

The interrelations between singular concepts and a theory, in which these concepts are embedded, are mutual. On the one side, a concept gets its own meaning by the relations of the theorems within this theory. On the other hand, if someone pursues a specific meaning of a concept, then a usual way is to develop a theory around this interpretation and see how far-reaching such a theory is. Thus, the mutual relations between a theory and specific concepts clarify also the concept. That applies also to the concept of probability.

Probability is open for many overlapping interpretations, which all show a complementary character. We will go into the details of an analytic clarification of the concept of probability as has been undertaken by the analytic scientist Stegmüller (1973) and by Hacking (1965). The usual way to determine the meaning of probability – from a modern perspective – is to build an axiomatic theory that reflects in its axioms central properties of probability and provides a rich system of theorems, which correspond to intuitive pre-conceptions on probability and allow extrapolating properties of probability beyond the pre-existing knowledge. Because of the mutual relation between probability and statistical inference – one cannot separate the two perspectives without loss of meaning – the meaning of probability cannot be determined in probability theory by itself. Especially, as there are various axiomatic theories of probability that would justify completely different perceptions. Thus, for investigating the meaning of probability, it is necessary to investigate also the way, in which we can build and justify statistical inference.

The philosophical debate in the 1930s has revived the conflict between material objective interpretations of probability, mainly related to an interpretation as something connected to relative frequencies and as a degree of belief that has significant closeness to something connected to personal probabilities. Both interpretations have emerged in an axiomatic theory, which serves as a justification of the respective interpretation, Kolmogorov (1933), for the relative-frequencies conception, and de Finetti (1937), for the degree-of-belief conception. Within the frame of probability theory, there was and

is a strong preference for the objectivist variant, especially for the need of an *empirical* view in physics where probability plays a lead role with the progress in thermodynamics by Boltzmann since the 1870s (see Steinbring, 1980).

However, the Bayesian Controversy in the foundations of probability looked beyond probability theory. If probability is something like relative frequencies in experiments, then it is the role of statistical inference to determine the conditions, under which we can measure probability by relative frequencies. Clarifying these conditions underlying the measuring process and the final precision attained, influences the meaning of probability. Thus, the methods of statistical inference shape the concept and the scope of probability. As these methods are based on probability, a mutual interrelation between probability and statistical inference emerges, which attains the character of a complementarity. The controversy was fuelled by the way, either school of probability tried to develop a theory of statistical inference. We go into more details of the controversy and clarify how the integration of statistical inference introduces a shift in rationality, objectivity, and scientific character of the methods.

Just to mention already here the resulting dilemma of the analysis: For the prevalent objectivistic perception of probability based on the meaning of probability as connected to something like relative frequencies, the theory of probability based on the Kolmogorov axioms (1933) is undisputed and justifies a frequentist perception of probability. Yet, the "theory" of statistical inference (which is not an axiomatic extension of probability theory) has severe rationality gaps, which can lead to bad decisions. For the subjectivist meaning in the form of degree of belief, the de Finetti (1937) theory of probability (also an axiomatic theory) justifies the perception of a personal belief. The advantage here is that the theory of statistical inference can directly be connected to conditional probability and thus to probability so that it is undisputed and rational. Yet, the personal meaning of probability based on the approach seems unacceptable for many scientists, especially from the point of view of physics where probability plays an eminent role.

2 Conceptions of Probability

While it was comparably easy to find a satisfactory solution for an axiomatic basis for probability, the crux was to build up from that framework a theory of inference that would allow for a connection of this conception to the real world. A statistical theory of inference that would also justify the tight connection between probability and statistical inference within an unambiguously accepted theory.

2.1 Emergence of probability

The great dividing line between several approaches towards probability is whether we perceive probability as a mere property of the real world, irrespective of beliefs and convictions of persons, or whether we perceive probability primarily as a degree of belief of a (rational) person. The first group of probability perceptions is called objectivist, the second subjectivist (as distinct from subjective, as we assume a rational person in the background).

Hacking (1975) addresses these aspects as the statistical and the epistemic side of probability. It is interesting to trace both sides in the early history. In the poem de Vetula from the 13th century (see DAVID, 1962), at first sight the epistemic view is pre-dominant, counting the possibilities to give reasons for or against a bet on A. Yet, as a consequence, one would put wagers on A according to expected gains as "you will learn full well how great a gain or loss any of them is able to be" (BELLHOUSE, 2000, p. 135, cited from BATANERO, HENRY, & PARZYSZ, 2005, p. 20). The latter statement refers to the statistical (empirical side) of probability. The historical example highlights that right from the beginning of documentation both sides, epistemic and statistical, were present and they back up each other to convey the meaning of probability. It even goes further than simply relate a combinatorial probability to statistical frequencies as it already establishes a meaning of a tendency of the game to produce specific results, which comes close to the modern propensity conception of probability by Popper (1959). By his Law of Large Numbers (LLN), Bernoulli (1713) was the first to establish successfully the relation between the epistemic side of the multiplicity of cases favourable to a bet (event) and the empirical-statistical side. His "golden theorem" reveals a further complementary character of probability – combinatorial probability makes sense only if it links to statistical probability and, as then they had no way to define statistical probability, it made sense only if it was linked to combinatorial multiplicity. We use the term complementarity from physics from Bohr (1928), in the sense of Otte (1984): Two concepts are complementary, if it is not possible to separate them without severe loss of meaning.

The next step of conceptual progress was the dispute about Bayes' (1763) theorem: First, the status of the so-called prior probability links neither to combinatorial multiplicity nor to relative frequencies. These probabilities were in fact derived by the conception of a bet with equal stakes for the outcomes in the experimental device and were based on "complete ignorance". The result was equal probabilities for that what now is called prior probabilities in Bayes' theorem. The character of this probability was seemingly subjective, though in some logical sense. The argument about total ignorance was later used (misused) by Laplace (1812) to justify his first definition of probability based on equiprobability.

The developments of modern physics, especially in thermodynamics in the second half of the 19th century made it necessary to find a suitable mathematical basis for probability in physics (see STEINBRING, 1980) as David Hilbert (1900) expressed it in his famous agenda of the most important mathematical problems. A theory directly based on the idea of probability as idealised relative frequencies, as von Mises (1919) attempted, failed. The indirect characterisation of Kolmogorov (1933) by an axiomatic theory was acknowledged as the solution of a probability theory that allows for a frequentist interpretation of the concept. Shortly after, de Finetti (1937) found an axiomatic solution for the idea of probability as a degree of belief. More about the fascinating history of the concept of probability is in Kapadia and Borovcnik (1991), or in Borovcnik and Kapadia (2014).

2.2 Classification of the meanings of probability

Before we expose the problematic of statistical inference, we summarise the conceptions of probability focussing only on three main meanings (the classical a priori, frequentist, and subjectivist), which are sufficient from an educational point of view. We refer to the terminology of APT, FQT and SJT of Çinlar (2011), and Borovcnik and Kapadia (2014).

Classical a priori theory (APT). The first definition of probability by Laplace (1812) of a combined event relates it to the ratio of favourable to all possible outcomes. It is well known that this definition is circular as it is based on the equal likelihood of elementary outcomes and likelihood is only another word for probability here. Yet, this conception of probability is essential not only for educational purpose. It determines values for the probability a priori, i.e., prior to any data, hence a priori theory.

Frequentist theory (FQT). The probability of an event is estimated from the observed relative frequency of that event in repeated trials. Exact values of probabilities are never obtained by this procedure. This is an a posteriori, an experimental approach based on information after actual trials have been done. The problematic here originates from the exactification of repeated trials. Von Mises (1919) used a so-called Regellosigkeit, in modern terms; this refers to independent identically distributed experiments (iid). Probability is defined as the limit of relative frequencies but such a limit can only be a façon de parler as in reality it never can be checked.

Subjectivist theory (SJT). Probabilities are evaluations of situations, which are inherent in the individual's mind and thus not properties of the external world, which is implicitly assumed in the two other approaches. The basic assumption is that individuals have their personal probabilities. These probabilities link to an implicit preference pattern between decisions. By data, one can update such personal probabilities by

Bayes' formula whence after enough data, an SJT would have nearly the same probabilities as an FQT. Yet, with less data, there remains also a distinction in the probability *values* not only in the *conception* of probability.

Structural view. The structural approach serves as a theoretical framework. Formal probability is implicitly defined by a system of axioms and the body of definitions and theorems, which may be deducted from these axioms. One can derive probabilities from other probabilities by means of mathematical theorems, yet, with no justification for their numerical values in any application. This structural approach does not intend to clarify the nature of probability, though its theorems are an indicator of possible interpretations. As the axiomatic theories for the two main conceptions of probability (FQT and SJT) correspond to each other, they are – structurally – quite similar.

In Table 1, we summarise the views of Barnett (1982, p. 65), Batanero, Chernoff, Engel, Lee, and Sánchez (2016), Good (1983, pp. 70), Borovcnik and Kapadia (2014, pp. 25), and for the scenario character of probability, Borovcnik (2006) (italics by the current author).

Some remarks on Table 1: Intuitive views may assume an archetypal character according to Batanero and Borovcnik (2016). In the sense of Popper (1959), propensity meaning reformulates a probability statement as a physical property of an object or situation to produce events, which may more pronouncedly called a tendency. Credibility differs from subjective probability by the way that the personal judgement is based solely on logical reasons (and not on personal preferences) though probability is attached to *judgements of "persons" about reality* rather than it were perceived as *property of reality*. Borovcnik and Kapadia (2014) lay a focus on personal probability as in decisive places in inverse inference (reasoning from data back to hypotheses) approaches of logical probabilities have failed.

Axiomatic view leaves the exact interpretation of probability open (HILBERT, 1900) though it builds a framework for an initial intuitive conception of probability; it was provided by Kolmogorov (1933) for the ideas of relative frequencies and by de Finetti (1937) for the idea of probability as subjective degree of belief. The scenario character of probability is resumed in modelling where one would base the analysis on a pseudo model, a scenario, regardless whether it fits well to the real situation or not. By the scenario approach, Borovcnik (2006) shows how it is possible to derive conclusions with a formal probability that has strong qualitative SJT character and definitely is void of an FQT interpretation.

Table 1: Classifications of meanings of probability in original sequence except for Good.

Barnett	Batanero et al	Good	Borovcnik and Kapadia
	Intuitive views: Just stating the role of intuitive views that may play a role for [understanding].	1) *Degree of belief* (intensity of conviction), belonging to a highly self-contradictory body of beliefs. [...]	
Classical: Symmetry considerations; 'equally likely outcomes'.	*Classical meaning*: Based on an assumption of equiprobability [...] justified by the disputed principle of insufficient reason.		*Classical a priori theory* (APT): Probabilities are given a priori (by symmetries or the principle of insufficient reason).
Frequency: Frequentist; empirical; relative frequencies in 'repeatable' situations.	*Frequentist meaning*: Based on a "limit" of relative frequencies in a repeatable experiment.	5) *Physical probability* (material probability, chance, propensity; this last name was suggested by K. R. Popper).	*Frequentist theory* (FQT): Probability is something like idealised frequencies. A posteriori estimated from data.
	Propensity meaning: Physical disposition or tendency, [to connect] long-run frequencies and an application to single cases [...].		

Table 1: Continuation

Barnett	Batanero et al	Good	Borovcnik and Kapadia
Logical: Objective; intrinsic 'degree-of-belief' as a logical measure of implication.	*Logical meaning*: Objective degree of belief, revised under new experience.	4) *Credibility* (logical probability, impersonal, objective, or legitimate intensity of conviction).	
Subjective: Personalistic; individual assessment of 'rational' or 'coherent' behaviour.	*Subjective meaning*: Subjective degree of belief, revised under experience.	2) *Subjective probability* (personal probability, intuitive probability, credence). [Consistency is required]. 3) *Multisubjective probability* [...].	*Subjectivist theory* (SJT): Probabilities are personal yet rational evaluations.
	Axiomatic view: Mathematical theory based on axioms.	6) *Tautological probability.* In modern statistics, it is customary to talk about ideal propositions known as "simple statistical hypotheses".	*Structural view*: Rich theory based on axioms that captures the basic idea of the concept.
			Scenario character of probability: Investigation on a "what if?" basis.

3 Characterisation of probability

In the analytic theory of science, a meaning of a concept is justified by developing a theory around this concept. The richer this theory is and the more phenomena of reality are reflected by derived concepts and theorems, the better is the justification for the initial meaning of the concept (STEGMÜLLER, 1973). Such a theory provides solutions for the phenomena; the interface between this theory and applications is characterised by this basic meaning, which allows translating from the situation in reality to the model at theory level. There are several criteria to judge the quality of a foundation: Does the theory allow dealing with as many phenomena and problems as possible? Is the development of the theory self-contained without rationality gaps? The best way to answer the second question is to formulate the basic rules for the concept in the form of axioms and then use only logic and mathematical relations to build the theory.

If we compare the two main ways to conceive the concept of probability, it becomes clear that there is some substantial interest to favour an FQT over an SJT interpretation as this would pave the way for a direct empirical connection for probabilistic hypotheses. That means that theories and statements (hypotheses) would have an empirical meaning and could be open for something like a statistical test that would provide an objective test of such hypotheses against empirical facts. That is what science should guarantee: statements that do not depend on personal judgement and "taste", but are rather open for such a test that is undisputed between different persons, i.e., free of subjective elements. If an explication of the concept of probability as connected to an FQT meaning is the goal of a scientific project in the foundations of probability, then the concept of probability cannot be solely explained by a theory of probability. The concept of probability then has to include proper methods for an empirical examination of probabilistic statements (statistical hypotheses). How rational such methods are will influence the conception of probability and its character as an objective term (i.e., independent of personal judgement).

FQT probability is indirectly determined by a theory, which partially characterises this concept. Though an axiomatic theory lets the concepts free of any meaning, the attempt of Kolmogorov is focused on a justification of FQT and usually probability within this approach is interpreted as something that is linked to relative frequencies in the long run. The term "partially characterised" refers to the fact that – in the objective perception of science – the empirical testability of probability has still to be established so that probability theory is incomplete. It has to be supplemented by a theory for statistical inference that provides methods for testing probability statements against empirical entities (the relative frequencies). In the objectivist framework, various methods have been developed for the purpose of statistical inference.

There are several axiomatic settings for probability so that there are 'legitimate' competitors for its meaning. Accordingly, *within* probability theory, it is not possible to restrict probability to one meaning. The openness towards diverse interpretations gets even more weight, as statistical inference lies *beyond* the scope of probability theory. Furthermore, the basic axioms do not cover the concept of independence, which is defined as a simple product rule for events. Yet, if the meaning of independence is determined by a definition solely based on probability, how can probability be explicated by the axioms and an additional reference to independence? Kolmogorov (1956, p. 9) recognises this vicious circle (italics from this author):

> "[...] one of the most important problems in the philosophy of the natural sciences is [...] *to make precise the premises* which would make it possible to regard any given real events as independent."

The usual way out of this dilemma is by explaining statistical independence by a vague reference to causal independence but that is more of a rhetoric trick than an explication as the two concepts are on completely different levels at different grades of precision (Borovcnik, 1984, p. 164). These difficulties with defining the concept of independence are one fundamental reason for adherents of an SJT conception of probability to replace it by the concept of exchangeability, which has been introduced by de Finetti (1937); see also Barnett (1982, p. 78). Exchangeability is a form of symmetry, which is easy to check by its influence on the betting behaviour: if the sequence of statements is of no influence on the elicitation of odds, then exchangeability applies. That is, to explain exchangeability needs no reference to vague causal arguments. In addition, the central theorems (such as the LLN) of probability still hold. Statistical independence also provides the basis for the concept of a random sample (which is necessary to combine several data).

Statistical tests of probability statements without independence (and hence without reference to random samples) would not be tractable. One has to guarantee independence for statistical tests, which is referred to a further background hypothesis that usually does not undergo the same statistical test but is rather inspected by "rules of thumb" or simply claimed to be valid. Thus, independence is constitutive for the concept of probability within probability theory and is a basic constituent of statistical tests so that the concept of independence substantially contributes to an FQT approach towards probability. Because of its inherent problems, subjectivists have replaced it by exchangeability.

4 Statistical and Bayesian inference

Statistical inference comprises methods within an objectivistic framework for probability related to an FQT interpretation of probability, while Bayesian inference subsumes methods within a subjectivist framework related to an SJT interpretation.

4.1 Developing a theory and methods of statistical inference

We start with some comments on hypotheses. We discern deterministic hypotheses, which are logical all-statements, and statistical hypotheses, which include a probability statement. A typical deterministic hypothesis is "All swans are white". We can falsify such a statement if we find one counter example (e.g., one black swan). A typical probabilistic hypothesis is "The probability of Head with a specific coin equals p".

The term "statistical data" comprises knowledge from observation (the empirical component) and background knowledge (the theoretical component). This background knowledge refers to the class of statistical hypotheses (as expressed by probability distributions), which form the basis of the empirical test (e.g., the family of normal distributions) if a specific statistical hypothesis (a normal distribution with specified parameters) undergoes such a test. Such knowledge attains the character of a background hypothesis (a hypothesis that is not questioned though it could be chosen differently). A further background *hypothesis* refers to the independence of single data; it allows regarding the data for the test as result of a random sample. The following example may illustrate the relevance of background hypotheses.

> The hypothesis $P(Heads) = p$ for a specific coin cannot be investigated in isolation, as it would seem at first sight. We have to assume that the number of Heads in n trials follows a binomial distribution $B(n, p)$; this type of distribution reflects the independence of the single trials. Within the test, this background hypothesis is not questioned.

Deterministic hypotheses (laws) are relative to acknowledged data definitely falsifiable but not verifiable. Yet, statistical hypotheses are neither verifiable nor falsifiable:

> If a sample of 20 throws yields 12 Sixes, it may seem convincing to reject the hypothesis "Probability of a Six in throwing a specific dice equals 1/6"; yet, there is no way to exclude logically that such an event can be observed if the hypothesis really applies.

There are differences in the character of hypotheses. As deterministic hypotheses can definitely be rejected (if one finds a counter example), only type-II errors can occur (i.e., a false hypothesis is erroneously not rejected). For statistical hypotheses, two types of errors can occur: additional to the type-II, also a type-I error can occur, when

a true hypothesis (it really applies) is erroneously rejected (BOROVCNIK, 2015). A deterministic hypothesis can be judged in isolation (if it is judged as false, it is definitely false) whereas it is not possible to judge a statistical hypothesis in isolation, as if it is judged as false, it still can apply. We need to know the degree to which it can apply – i.e., we need the type-II error (a wrong hypothesis is erroneously not rejected) but that depends on the alternative hypotheses considered. Consistently, checking a deterministic hypothesis has to provide rules for rejecting such a hypothesis, while a procedure for checking a statistical hypothesis can only provide statements about the support of hypotheses. Preferably, support functions would be probabilities but an objectivistic framework excludes that, as probabilities must have an empirical connection to relative frequencies in an experiment and there is no such random experiment for hypotheses. A probability structure for the support of hypotheses would clearly be an option for the subjectivist position, as herein probability is not restricted to empirically testable elements.

For the objectivist position, the used methods comprise the likelihood of statistical hypotheses and indirect approaches such as the Neyman-Pearson test policy (NEYMAN & PEARSON, 1928; NEYMAN, 1937; PEARSON, 1966) and the Fisherian significance test (with "fiducial probabilities"). These approaches use direct or indirect notions to express the support of hypotheses that lack a probability structure so that they are harder to interpret.

4.2 On the logic of support

We seek a solution for an approach for statistical inference first as an extension of an FQT probability (which is characterised by the Kolmogorov theory). We have already argued that this statistical inference has a direct influence on the character of probability. First, we define the conception of support of hypotheses by an example.

> We toss a coin 12 times. As a result, we get 9 Heads. We are interested in the probability p, with which the coin falls Head:
>
> $$P(\text{No of Heads} = 9|p) = (12 \text{ choose } 9) \times p^9 \times (1-p)^3.$$

This probability for the empirical result *No of Heads = 9* is much smaller under the hypothesis of an ideal coin ($p = 0.50$) than it is under the alternative hypothesis of ($p = 0.75$): the probabilities are 0.0537 and 0.2581. The hypothesis "the coin is ideal" provides a much smaller probability for the observed event than the alternative hypothesis does. Therefore, intuitively, we have to take this alternative hypothesis much more into account; we say, this alternative (hypothesis) is much better *supported*.

This is a typical likelihood argument: The likelihood L of a hypothesis h in view of a specific observation E is simply the probability that h provides for the observation:

$$L(h|E) := P(E|h).$$

There are some immediate statements about likelihood support arguments to place here. Likelihood arguments

- attain only sense if we compare their values between competing hypotheses;
- are NOT probability statements for hypotheses (which would be grossly inconsistent within an approach aimed at an experimentally testable conception of probability);
- miss to provide logical conclusions from (empirical) data to (theoretical) hypotheses.

The so-called likelihood principle of Birnbaum (1962) states that for the judgement of hypotheses it suffices to refer solely to the likelihood of hypotheses under the actual observation and nothing else. As rational as such a principle may seem, there is no further justification for it. It is worthy to note that not all standard tests in statistics follow the likelihood principle. Stegmüller (1973) characterises the concept of support by axioms and interprets support always as likelihood support. In order to make an efficient use of the concept of likelihood, Stegmüller introduces the term combined statistical statement as $h = \langle D, E\rangle$, where D is a statistical hypothesis that specifies a specific probability distribution, and he defines the likelihood L of combined statistical statements as

$$L(h) := P(E|D).$$

The difference to before is, that now both E and D could be fixed (earlier E was fixed only and D was perceived as variable). The "symmetric" form of likelihood serves for two purposes:

- The conclusion from data to hypotheses (inverse inference) if E is fixed.
- The conclusion from hypotheses to data (direct inference, or single case) if D is fixed and E is variable.

The so-called likelihood rule determines the shift from likelihood values to support statements. Hacking (1965) presents the likelihood rule in a very general form, while Stegmüller (1973) modifies the rule to a logical conjunction of the likelihood principle and a rule for the single case. The generalised likelihood serves its purpose as it combines

both basic statistical tasks (see before). A more detailed representation may be seen from Borovcnik (1984, pp. 200-207). Stegmüller (1973) follows the ideas exposed in Hacking (1965, pp. 59).

4.3 The project of characterising probability on an objectivist basis

With the likelihood principle and the generalised likelihood function, the project of an explication of the concept of probability based on Kolmogorov's theory and an FQT interpretation is finished. The clarification of probability from an analytic point of view is complete and Stegmüller perceives statistical probability in this extended sense (BOROVCNIK, 1984, p. 179). We are going to represent the critique of Stegmüller and Hacking to other approaches to a statistical inference based on an FQT interpretation of probability and the superiority of the likelihood test theory. We insert some meta-considerations about the type of this project and elaborate further on direct and inverse inference.

We may see the procedure of Stegmüller (1973) in the following way. He starts from a pre-fixed idea of probability (it has something to do with relative frequencies in the long run); he develops an analytic explication of the concept of probability by a reference to Kolmogorov's theory and builds a theory of statistical support based on the likelihood principle and the (generalised) likelihood rule. From that basis, it remains to develop statistical methods for the two basic statistical tasks, direct and inverse inference.

Before we go into the details of statistical methods, we already present an essential weakness of Stegmüller's approach here. By his symmetric generalised likelihood, he succeeds to solve both direct and inverse inference in the same formal way. Yet, the symmetric approach levels off the essential differences between the two problems. The main reason for that undue symmetry is that the single case (direct inference) bears the structure of probability while inverse inference lacks such a structure with probabilities. Therefore, it is an unsuitable way to wipe out the differences and declare the problem as solved. Since Bayes (1763), there has been a dispute about the status of the prior probabilities used for the inverse conclusion. While the objectivist conception of probability with an FQT meaning has proven successful for the single case, inverse inference causes a shift from objective to subjectivist perceptions. Consistently, we will investigate objectivist methods for statistical inference developed so far and we will use the results of the analysis for a re-evaluation of the subjectivist-objectivist controversy.

In that re-evaluation, the problematics of background hypotheses plays a special role. There are two different types of background hypotheses: i. A family of distributions,

on which the mathematical considerations (the mathematical model) is based upon. ii. The independence of repeated trials that lead to combined empirical data. Of course, considerations of simplicity decisively influence the determination of background hypotheses. In that respect, statistics makes no difference to other areas of mathematics. Yet, Stegmüller (1973, p. 135) explicitly claims that it is possible to apply the statistical methods based on his theoretical framework to these background hypotheses with no further difference to the specific statistical tests. That point, however, remains a mere claim, as we will see below.

4.4 Direct inference and inverse inference

We identify two distinct statistical tasks, the conclusion from a specific probability distribution (a statistical hypothesis) to 'future' data, and the conclusion from observed data to potential statistical hypotheses.

Direct inference or single case. What does a probability statement mean for reality? More precisely, what can we conclude from a statistical hypothesis (a probability distribution) for a 'future' event? The single-case rule determines the meaning of such probability statements: The support for the statistical hypothesis e "E occurs in a concrete trial" is exclusively determined by $P(E)$, which is the objectivist probability for the event E as specified by a model. There are several attempts to justify the single-case rule (see BOROVCNIK, 1984, pp. 194-199):

Naïve objectivists-frequentists justify the single-case rule by "long-run" arguments or they request "axiomatically" that an "optimal" decision in the long run is also optimal for the single case without any deeper reason. Yet, considerations in Borovcnik (2015) show that optimal decisions adapted to the single case are different from those that optimise long-term behaviour, which highlights that it is not easy to mediate between single-case and long-run.

On a theoretical basis, Stegmüller justifies the rule by a reference to the likelihood support theory. This approach, however, is circular: On the one side, the likelihood rule attains its authorisation only because it provides a derivation of the single-case rule; on the other side, it is only possible to justify the single-case rule by a reference to the generalised likelihood rule.

The propensity meaning of probability, suggested by Popper (1959), demands the applicability of a probability statement to a single case by a "tendency" argument; yet again, it offers no justification, as it is a rewording rather than an elsewhere grounded property of probability.

A subjectivist justification refers to accepted odds in bets and if no further information becomes available for the single case, then the odds are completely determined by

the probability so that the probability statement in general and the statement for the single case coincide.

Yet, direct inference really is a minor part of the controversy between objectivists and subjectivists. It is worthy to note that one has to separate the problem of the justification of the single-case rule from the problem of the correct application of this rule, as the probability statement could be ambiguous. This causes difficulties regardless whether one adheres to an objectivist or a subjectivist position towards probability.

Inverse inference – Inference from data to hypotheses. Inverse inference means to conclude from empirical data to probability models latent behind the random process. Bayes was faced with the problem of a Bernoulli chain producing binary data with an unknown probability p for the *1s (1 – p for the 0s)* and wanted to make an inference about p when a series of n trials results in *k 1s (and n–k 0s).* Bayes used an argument – known as Bayes' postulate – with complete lack of knowledge about p to derive equal probabilities (a uniform distribution) for the possibilities, which are here continuous values of p in the interval $(0, 1)$. In Bayes' setting, the parameter p in the random experiment provides a binomial distribution for the data (the number of 1s). Inverse inference on p relative to the data k of n is based on a "revised" distribution for p (a beta distribution). For the details, the reader may consult Good (1983) or Barnett (1982). We are only interested in the "structure" of the situation, which can be summarised as discrete version of Bayes' theorem; a simplified case is in Batanero and Borovcnik (2016), the general case with continuous densities is in Stegmüller (1973, p. 121).

Bayes' formula. Let $H_1, H_2, \cdots, H_r$ be r exclusive and exhaustive hypotheses for a stochastic situation and E be an empirical observation. If the probabilities $P(H_i)$ were known as well as the probabilities for E under the various hypotheses, i.e., $P(E|H_i)$, then the hypotheses attain a new probability conditional to E in the following form:

$$P(H_i|E) = P(H_i) \times P(E|H_i)/P(E) \; \alpha \; P(H_i) \times P(E|H_i). \tag{4.1}$$

Hereby, the term $P(E|H_i)$ represents the likelihood of H_i relative to the empirical data E, and $P(H_i)$ stands for the prior probability (prior information, an information, which applies before the data E becomes known). The term on the left side is called the posterior probability of the hypothesis relative to the data E. This posterior probability attains the structure of a probability (i.e., it fulfils Kolmogorov's axioms, or better, de Finetti's axioms) and integrates the data into the support function. (The last part of the formula after the α sign is a simplification. As long as the data E is fixed, the solution is *proportional* to the numerator and the denominator works only as a normalising constant to make the sum of all probabilities to 1.)

We mentioned that Bayes – within his experimental device – had derived a uniform distribution for the parameter p. Laplace (1812) extended his argument to the principle

of insufficient reason, which obviously represents an epistemic flaw. How can the principle of insufficient reason, overly interpreted as absolute no knowledge, emerge into a uniform distribution on the possible cases, which undisputedly represents some sort of knowledge.

Bayes' formula bears a great potential. Together with a so-called conditional conditionalisation (SEIDENFELD, 1979, p. 5-8; BOROVCNIK, 1984, p. 211), it forms a *complete inductive logic*. However, the price for that achievement is very high indeed, as the information about the hypotheses on the prior probabilities is usually not frequentist so it leads beyond the objectivist position in the foundations of probability. Objective Bayesians, as Jeffreys (1948), tried to find logical justifications for Bayes' postulate. This endeavour led to irresolvable problems that we may circumscribe by violations of the invariance principle. Of course, such prior probabilities usually are not open to an experimental control by relative frequencies. Overall, objectivists, starting from Venn (1866), refuted Bayesian inference as a solution for inverse inference as they request that the constituents of methods for judging statistical hypotheses have to link to objectivist probabilities only (i.e., probabilities that allow for an FQT meaning). Modern subjectivists as Savage (1962) circumvent the justification of the uniform distribution (in Bayes' postulate) by grounding probability as degree of confidence. The objectivist position discredits such "solutions", however, as private and thus non-scientific hypotheses testing. Yet, one still has to check how the various approaches of objectivists cope with inverse inference. This is the target of the following section.

5 Objectivist test theories

Within the objectivist school, a strong position in favour of the likelihood test and against the NP approach has emerged. Likelihood tests are based on the likelihood support theory, which is thought to be constitutive for an objectivist conception of probability. To the contrary, Neyman and Pearson (1928) circumvent the support logic by their policy of testing hypotheses repeatedly. The Fisherian significance test (1925) – though intuitive – is no acceptable solution from the perspective of foundations as it lacks any consideration of alternative hypotheses.

5.1 Likelihood test theory

In the literature on the foundations of probability and statistics, the likelihood test theory plays a prominent role since the formulation of the likelihood principle by Birnbaum (1962). Consistently, Hacking (1965) and Stegmüller (1973) favour this test theory in

their discussion on the controversy in the foundations. Central notion of this approach is the likelihood support.

Early discussions of Bayes' theorem have revealed enormous problems in finding an objective justification for the prior probabilities $P(H_i)$. It was even doubted – not only by Venn (1866) – that it would make sense to attribute a probability to hypotheses in an objectivist sense as it is not open for an empirical test in an experiment with relative frequencies. Thus, what was left of Bayes' theorem was the term $P(E|H_i)$, the probability of an empirical event given the hypothesis H_i, which was interpreted as support for H_i in the sense of the higher the probability for E, the *higher the support for* H_i. Such a thinking formed apparently the background of arguments such as by Arbuthnot who interpreted the fact E (80 years with a majority of boys in birth statistics) as an argument for the hypothesis H (a divine order) as $P(E|H^c)$ is very small. Arbuthnot interpreted the complement H^c to the divine order H as "randomness".

In the scientific analysis of the likelihood $L(H|E) = P(E|H)$, it became soon clear that the likelihood cannot be interpreted in absolute values but only compared to the likelihood of other hypotheses (HACKING, 1965, p. 59). Birnbaum (1962, p. 271) formulated the *likelihood principle* according to which nothing else apart from the likelihood function should influence the support of hypotheses. With this likelihood principle and a translation of likelihood values into the support of hypotheses, a test theory could be justified. The support logic thus is characterised as a comparison between several hypotheses. Loosely speaking, of two hypotheses, we reject the null hypothesis h_0 if the alternative hypothesis h_A has a support that is larger than a specified factor in comparison to the support of h_0, which is summarised in the following two definitions. The support here is operationalised by the likelihood function. A likelihood test to the level γ based on the empirical data E rejects the hypothesis h_0 in favour of h_A if and only if

$$L(h_A|E)/L(h_0|E) \geq \gamma. \tag{5.1}$$

If the alternative hypothesis is composed of several distributions (hypotheses), which we denote as a set Δ, then we have to take the supremum (a mathematical refinement of the maximum) of the single likelihood values, i.e.,

$$\sup_{h_A \in \Delta} L(h_A|E)/L(h_0|E) \geq \gamma. \tag{5.2}$$

- The choice of the critical number γ quantifies the degree of *better support* and determines the critical region (all data that lead to the rejection of the hypothesis h_0).

- Relative support between two hypotheses is difficult to interpret. For example, what does it mean that the alternative has double or more support than the null hypothesis?

It is remarkable that the likelihood support is not defined symmetrically between the hypotheses as – from the point of support – there is no need to attribute the role of a null hypothesis to one of the hypotheses under scrutiny (for more details, see BOROVCNIK, 1984, p. 258). While the likelihood test theory is also favoured in a theoretical framework of statistics and higher-quality applications, and while it is that test theory, for which the justifications from an analytical point of view are the best, for the practice, it remains the difficult task to choose the value of γ of a test to apply. There is hardly a guidance from its interpretation though γ finally decides about the decision between the hypotheses. Therefore, there is still a need for a test theory that allows a more transparent choice of the indices of a test that determines the actual decision. This is the case for the Neyman-Pearson test theory, which we discuss subsequently.

5.2 Neyman-Pearson test theory

This test theory represents that one that is still dominating statistical practice and the teaching of statistics. Neyman and Pearson (1928) follow a completely different path in their approach. They do not base their considerations on the support of hypotheses but direct their criteria for finding an optimal test to the general properties of the test method rather than on properties of the hypotheses. They operationalise the goodness of a statistical test by an explicit interpretation of probability in an FQT sense.

For direct inference, their way is characterised by a mixture of equiprobability and a frequentist meaning. Already Fisher (1935) criticised their justification fundamentally; yet, we will not delve into the details, as anyway direct inference is not the core of the controversy between objectivists and subjectivists.

For inverse inference, they offer their policy of hypothesis testing, which is decision oriented. Rather than aiming at a support of the hypotheses under scrutiny including the actual data, they determine a concrete test by general properties of the "decision rule" that are operationalised as FQT probabilities. In the simplest case, if the null and the alternative hypotheses fix a single probability distribution, they compare the scenarios derived from these distributions on the space of all samples. Any decision rule is characterised by a subset of samples, which lead to the rejection of the null; this is called *rejection region R*. Neyman and Pearson (1928) determine the conditional probabilities of the already discussed type-I and type-II errors.

If we denote a rejection dependent on the result of a sample as T_A (sample lies in R) and a non-rejection as T_0 (sample lies in the complement of R), we have:

> α = type-I error of the decision rule = $P(T_A|h_0)$ = probability of erroneously rejecting a valid h_0.
>
> β = type-II error of the decision rule = $P(T_0|h_A)$ = probability of erroneously staying with a wrong hypothesis.

For simple hypotheses (each fixes one distribution), the errors have a probability value, which is interpreted in an FQT sense as relative frequency of erroneous decisions in the long run under the two scenarios that are described by the null and the alternative hypothesis. The complement of the type-II error, $P(T_A|h_A) = 1–\beta$ can be interpreted as *power* of the test to "detect" an alternative h_A if it in fact applies (is true). The α error is also termed as size of the test (not to be confused with size of a sample). Neyman and Pearson pursue the following strategy to find a test with optimal properties: First, the size of a test is pre-given. That restricts the considered tests, which we ideally just identify with the rejection region (the subset of all samples that lead to T_A, i.e., the rejection of the null hypothesis). Second, among all tests with size α, that test is optimal that maximises the power (that minimises the type-II error). It is essential that size and power of a test do not relate to the hypotheses under scrutiny but they are properties of the test procedure. Size and power of a test are interpreted as relative frequencies in the long run when the test is applied repeatedly.

In the extension of the test theory to hypotheses that are composed of several probability distributions, Neyman and Pearson introduce further criteria to guarantee the existence of unique optimal tests. Essentially, these criteria are unbiasedness and invariance. Despite all critique against the Neyman-Pearson test policy, one has to state that their optimal tests coincide with tests from the likelihood test theory for many standard problems. That means one may consider the values for α and β as useful to fix the ratio γ of relative support for the two hypotheses.

5.3 Critique against objectivist test theories

Within the objectivist school, there has emerged a strong position in favour of likelihood tests and against the Neyman-Pearson (NP) test policy. Likelihood tests are based on the likelihood support theory, which is thought to be constitutive for an objectivist conception of probability. To the contrary, Neyman and Pearson (1928) circumvent the support logic for hypotheses by their policy of testing hypotheses repeatedly. We reproduce the core of the critique of Hacking (1965) and Stegmüller (1973) against

the position of Neyman and Pearson; exactly the same examples, however, provide an opportunity to lead the likelihood test theory and the support logic beyond its boundaries of validity and plausibility.

Critique against the concepts of size and power. The following example of Hacking (1965, p. 87) is directed against the concept of power; at the same time, the example indicates a weakness of the NP approach, which may be termed as forward vs. backward look. Given the following test problem of h_0 against h_A and two tests R and S (see Table 2). It is worthy to note that these tests act somewhat complementary. Both have the same size of 0.01 but Test 1 has a much higher power. Yet, Test 1 has obviously a flaw as it stays with h_0 if E_1 occurs, an event, which has zero probability under h_0, i.e., this event cannot occur if h_0 applies! That means, Test 1 is much better than Test 2 in the forward look (before samples are available), yet in the backward look, it comes to an absurd decision if the sample yields E_1.

The critique focuses on an interesting aspect of tests, namely the forward and backward look, yet it is inappropriate. Why would one not prefer Test 3 with rejection region $S^* = \{E_1, E_3\}$, which also has size 0.01 but power of 0.98, which is the optimal test for size 0.01. NP *optimise* the power for a given size of tests, so that one cannot criticise the concept of power in isolation as it is essential to optimise power, which interrelates the concepts of power and size.

Table 2: Hypotheses and outcomes in Hacking's example criticising power

Hypothesis	$P(E_1)$	$P(E_2)$	$P(E_3)$	$P(E_4)$
h_0	0.00	0.01	0.01	0.98
h_A	0.01	0.01	0.97	0.01

Source: Hacking (1965, p. 87)

Test 1 with rejection region $R = \{E_3\}$ has a size $\alpha_1 = 0.01$ and a power $1–\beta_1 = 0.97$.
Test 2 with rejection region $S = \{E_1, E_2\}$ has a size $\alpha_2 = 0.01$ and a power $1–\beta_2 = 0.02$.

Critique against Neyman's test policy. Neyman states that it is impossible to get knowledge about the specific hypothesis under scrutiny. Rather, one has to find rules for rational decisions, which guarantee that erroneous decisions do not exceed some levels in the long run. This is in-line with Neyman's naïve frequentist views. It may suggest some plausibility for his test policy but is not sufficient for an analytic justification of his test policy.

The probabilities for type-I and type-II errors are *conditional* (!) probabilities relative to different scenarios and they are related to repeating the test situation, which amounts to a meta experiment. Repeating the test situation would imply – if the relative frequencies should relate to probabilities – that the test situation is repeatable under

the same conditions independently (the iid assumptions, independent and identically distributed). This may be a proper assumption about the process of random sampling for the data on which we base the test. Yet, it is doubtful whether we can think of the modelling of the whole test situation as an iid situation (including the variables investigated, the distributions assumed, the hypotheses formulated, and the test statistic chosen; even the initial problem would have to be always the same).

Other modellers would end up with different variables and a different model. The chosen model always implies a modelling act with possible errors and this process of modelling cannot be reproduced ad infinitum. "Essentially, all models are wrong, but some are useful", as Box states (BOX & DRAPER, 1987, p. 424). As different teams would obtain different models, a direct replication of the "subjective" model as an iid process establishes an artefact. The idea of a repeated test situation under exactly the same probabilistic conditions may be appropriate in quality control, the context, in which Neyman and Pearson developed their test policy. However, also here it becomes essential that one needs to have a prior information about the frequency of h_0 and h_A, in order to get a long-run frequency of erroneous decisions as the type-I and II errors are only conditional to distinct scenarios and have no overall (unconditional) meaning.

Forward and backward look. A further example of Hacking (1965, p. 89) should corroborate the superiority of the likelihood test theory.

> A random experiment may yield outcomes $0, 1, 2, \cdots, 37, \cdots, 99, 100$. In the following, we refer to 37 as one specific outcome, which plays the role of a representative. The task is to test the simple null hypothesis h_0 against a compound alternative $h_A = \{j_1, \cdots, j_{100}\}$. The single distributions are visible from the lines of Table 3. The test should have a size of $\alpha = 0.10$ with a sample size of 1.

Table 3: Hypotheses and outcomes in Hacking's example against NP test theory

Hypothesis		$P(0)$	$P(1)$	$P(2)$	$\cdots$	$P(37)$	$\cdots$	$P(99)$	$P(100)$
h_0		0.900	0.001	0.001	$\cdots$	0.001	$\cdots$	0.001	0.001
	j_1	0.910	**0.090**	0.000	$\cdots$	0.000	$\cdots$	0.000	0.000
	j_2	0.910	0.000	**0.090**	$\cdots$	0.000	$\cdots$	0.000	0.000
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
h_A	j_{37}	0.910	0.000	0.000	$\cdots$	**0.090**	$\cdots$	0.000	0.000
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	j_{99}	0.910	0.000	0.000	$\cdots$	0.000	$\cdots$	**0.090**	0.000
	j_{100}	0.910	0.000	0.000	$\cdots$	0.000	$\cdots$	0.000	**0.090**

Source: Hacking (1965, p. 89)

We first discuss the likelihood test:

- If 0 occurs in the sample, then h_0 has a support of 0.900 and each of the single distributions of h_A has a support of 0.910 whence the support of h_A is slightly better than for h_0. That means 0 speaks slightly for h_A; the ratio is 0.910/0.900 in favour of h_A.
- If any number different from 0 occurs (imagine it is 37), then h_0 has a support of 0.001 while j_{37} has a support of 0.090 and hence the alternative has (as supremum of all likelihoods over the alternative) a support of 0.090. That means, any result different from 0 has a likelihood ratio of 90 in favour of the alternative.

With a critical value for γ of say 4, we decide for h_0 if 0 occurs and for h_A if any other number occurs. Stegmüller (1973, p. 183) speaks of an intuitive plausibility of the support argument, which corroborates the likelihood test as very good. More complicated is the situation for the NP optimal test:

> The NP test cannot use $\{1, \cdots, 100\}$ as rejection region though this would fulfil the requirement for the size $\alpha = 0.10$ but the power would be $1–\beta = 0.090$ for any j_k from the alternative so that the power would be less than the size and the test would be *biased*. To repair the situation, the NP test has to use the additional criteria of unbiasedness and invariance to come to a best test that decides in the following way:
>
> Reject h_0 if the result of the random experiment yields 0 and an additional random experiment (that has nothing to do with the present hypotheses) is performed and ends with a specified result R that has a probability of 1/9. Otherwise, do not reject h_0. Such a randomised test has a size of $0.900 \times 1/9 = 0.100$, which fulfils the requirements.

Hacking and Stegmüller (1973, p. 183) summarise their critique against the NP optimal test and argue in favour of the likelihood test:

- The likelihood test is based on intuitively plausible considerations of support of the hypotheses after the sample when the empirical data is known.
- The NP test is derived by formal criteria, which should guarantee a good behaviour of the test before the data is collected. Rational behaviour in the forward look can be quite irrational in the backward look.

We acknowledge that the critique reveals essential differences between the two competing test theories. Yet, the critique is weak as its scope of validity is very restricted. It addresses low-ranked criteria of the NP approach and it breaks down with samples

with more than one data. An example of Seidenfeld (1979, pp. 51) backs the critique, yet it is formulated in terms of confidence intervals so that we omit it here.

Critique against the likelihood theory. We are going to undermine the plausibility of the likelihood test. For that purpose, we use a formal prior distribution on the single hypotheses as equal probabilities. This prior distribution is not arbitrary as it reproduces essential properties of the likelihood test from before (the 90 times higher support for the alternative if the experiment shows any number different from 0). Hacking's example is continued and Borovcnik (1984, pp. 258) uses it against the likelihood test as it reveals severe deficiencies of the likelihood test.

> For the formal prior distribution on the hypotheses with $P(h_0) = 1/101, P(j_k) = 1/101$ for $k = 1, 2, \cdots, 100$, we obtain the following posterior probabilities using Bayes theorem (we use 37 as a representative for an outcome different from 0):
>
> $$P(h_0|37) = 1/91 \qquad P(h_0|0) = 9/919$$
> $$P(h_A|37) = 90/91 \quad P(h_A|0) = 910/919\ .$$

These posterior probabilities take over the role of the support function $L(h|x) = P(h|x)$. We can state that any non-zero data supports the alternative by a factor of 90, which is exactly what the likelihood test from before is based upon. Yet, we also can see that the alternative hypothesis is supported even much better by the result 0 as the factor in favour of the alternative is larger than 100. That means, 0 speaks much more (in the sense of support) for the alternative than 37 (or any other non-zero outcome). Thus, for a test based on support, rejection should focus on outcomes 0 rather than on any non-zero outcome (that is what the NP test in fact did!).

Consequently, it is by no means plausible to accept that hypothesis that has a better support; in particular, support statements are void of plausibility and may lead astray. The likelihood test bases its intuitive "superiority" on the fact that better support means something positive. Yet, with a result of say 37, it draws the high likelihood ratio of 90 in favour of the alternative in an unjustified way. It assumes that we have waited for the 37 to occur. Any other number would have the same effect. We see that better support is not something that is an unquestionably good property.

There may be two implications from this critique: i. One should carefully combine both competing objectivist test theories, as none of them proves superior. ii. One should leave the objectivist position behind and specify prior distributions on the hypotheses under scrutiny. Stegmüller (1973, p. 293) gets to the limits of the likelihood test theory via the paradox of Kerridge; yet, he does not abandon the objectivist position:

> "Perhaps the most convincing thing that remains is the concern: [...] the necessity of a radical subjectivation of the natural sciences was pointed out if the [subjectivist] view of the concept of probability were to prevail. The science theorist, before he swallows this bitter pill, will look for another solution."

In conclusion, Stegmüller declares Fisher's fiducial argument as a potential way out of the crisis. This argument, however, faces the same difficulties as the programme of the Objective Bayesians (e.g., JEFFREYS, 1948), which may be seen from Good (1971). Thus, all attempts to avoid prior probabilities in Bayes' theorem or to find an objective justification for specific priors are doomed to irreparable inconsistencies.

6 Significance test and testing background hypotheses

We show the problematic of the Fisherian significance test so that it misses to count as an alternative for the likelihood test. There was already a problem with tests for background hypotheses, which can be tested only by significance tests. This completes the critique against objectivist test theories as the background hypotheses have to be checked by other methods rather than statistical tests so that a further rationality gap arises that finally shows the theoretical weakness of an explication of probability on a closed objectivist basis with an FQT meaning.

6.1 Significance test – measure of discrepancy

Gigerenzer (2004) identifies early traces of statistical inference in the first significance test by Arbuthnot (1710) in his *proof* of a divine order of gender. Arbuthnot's argument is cited from Borovcnik and Kapadia (2014, p. 18)

> "The probability is very small that for 80 successive years more males than females are born, so the hypothesis that the gender proportion is equal has to be rejected as it would produce an observation with a probability of $(1/2)^{80} \approx 10^{-25}$. Therefore, the alternative hypothesis must hold (probability for boys greater than for girls), which Arbuthnot interpreted as an expression of a divine order."

Once we decide the limit for moral probability, any hypothesis that attributes to the observed event a probability lower than this limit, is considered to be probabilistically *disproved*. Hereby, moral probability means the idea to neglect all probabilities lower than this threshold. This idea has occurred at several places in the emergence of probability. For example, Borel (1943) argued to equate small probabilities to zero:

for a human being, the threshold is 10^{-6}, in the history of earth, it is 10^{-15} and, for the cosmos, it is 10^{-50}. Such arguments illustrate how tight the concept of probability connects to considerations of statistical inference.

There are two perceptions about the significance test, which has been introduced by R. A. Fisher (1925): i. The significance test is a degenerate NP test, as it does not take into account alternative hypotheses. ii. The significance test provides a decision scheme analogous to the indirect proof in mathematics. Rather than providing a counter example that disproves the null hypothesis, the observation is considered as a relative, a statistical "disprove" of the null if it has a small significance level α (later referred to as p value). In the Fisherian tradition, the significance level α was interpreted as a discrepancy measure. If the probability α of the "observation" conditional to a hypothesis h_0 is small,

$$P(\text{"observation"}|h_0) = \alpha, \tag{6.1}$$

then this is a statistical argument against the null hypothesis. The significance level was ordinally scaled and void of a frequentist interpretation. For a small number α, the result of the decision to reject the null can now be interpreted in the following way (α was usually compared to the thresholds 0.05, 0.01. 0.001): Either a "rare" event has occurred (as a qualitative statement) or h_0 is wrong. Fisher himself did not pre-specify α before the data is gathered. He interpreted the value of α as a direct property of the hypothesis under scrutiny. As an NP test, the significance level can be interpreted in the long run; yet, it is a degenerate NP test as it has been derived without reference to alternative hypotheses whence Neyman's rationality principle (to maximise the power among those tests that have a common, pre-fixed size) cannot be applied.

We give an example for illustrating the two approaches.

Example 6.1. If two variables are jointly normally distributed (the background hypothesis), then we can find an optimal NP test for the null hypothesis h_0 of correlation $= 0$ (which then is equivalent to the independence of the two variables) against the alternative hypothesis of correlation not equal to 0. For this problem, it is possible to find an NP test with size $\alpha = 0.05$. This setting allows also for a Fisherian significance test. If an observation occurs that lies in the rejection region of the NP test, then the Fisherian significance level would be less than 0.05 so that we might come to the conclusion to reject the null hypothesis (correlation $= 0$).

Yet, in the Fisherian framework, one would not pre-specify the 0.05, as is done in the NP sense where the value for the size has to be fixed before the data is gathered. In both cases, the test can be perceived as test of independence against dependence (within the background hypothesis). Consider the following variant of the situation.

Example 6.2. The background hypothesis of the joint normal distribution is now missing. Yet, we can at least assume the data to stem from a random sample of the joint distribution. We have to test the null hypothesis of no correlation. As there is no way to find and order all hypotheses about the dependence of the two variables (and the consequences upon the correlation coefficient between them), there can be no NP optimal test as we cannot optimise the rejection region to maximise the power. Yet, in the Fisherian framework, we can just reorder the data (permute the y values as they are attached to an x value). If the variables are independent, then any reordering has the same justification. From all possible ways to reorder, we calculate the correlation coefficient so that we have a theoretical distribution of the correlation coefficient under the null hypothesis of independence. From this null distribution, it is possible to determine the significance level of the initial data and base a Fisher test on it.

There has been a fierce debate between Fisher and Neyman about their approaches. Fisher criticised Neyman's primitive FQT views about probability, while Neyman denied the rationality of the significance test as it lacks a possibility to optimise the test against alternative hypotheses. Fisher insisted that his discrepancy measure (later called p value) provides a valid support argument for the null hypothesis under scrutiny. Consistently, Fisher elaborated his ideas on fiducial probabilities, for which he was fiercely criticised by the community of statisticians (see SAVAGE, 1976). Recently, there are endeavours to revive his approach of fiducial probabilities, which are some kind of substitute for "prior" probabilities for hypotheses yet explicitly avoiding to use an SJT meaning. If in the applications of statistics a likelihood approach does not lead to satisfactory solutions, a careful analysis of the problem would be tried using a prior distribution, which is formally adapted to the problem under scrutiny (in the sense of a mathematical model and not as an expression of an SJT probability a priori).

From a perspective of analytical science, the significance test does not count as a suitable method for testing hypotheses as it neglects alternative hypotheses. This has a severe consequence upon the way to test background hypotheses about the type of distribution, or the independence of data. The critique against the method is widespread: "The illogic of statistical inference" (GUTTMAN, 1985), "A review of an old and continuing controversy" (NICKERSON, 2000), "An investigation of the false discovery rate and the misinterpretation of p values" (COLQUHOUN, 2014), are only a few examples that illustrate the critique against the significance test. Many voices ask for a ban or a replacement of significance tests (HUNTER, 1997; MULAIK, RAJU, & HARSHMAN, 1997; GORARD & WHITE, 2017). Lane (1980) summarises the controversy about the nature of Fisher's fiducial probability and Jeffrey's logical probability – both intended to rescue an objective meaning for the inference based on Bayes' theorem. Seidenfeld (1979) concludes that reconstructions of the fiducial argument lead either to a restric-

ted validity or to logical inconsistencies. In this way, the fiducial programme ends up with the same problems as Objective Bayesians (JEFFREYS, 1948), namely that it is not possible to represent total ignorance on prior probabilities in Bayes' theorem by statistical information.

6.2 "Tests" for background hypotheses

In any statistical test, both the family of distributions that describe the generation of the data and the independence of single data are crucial pre-assumptions. All further test results seemingly depend on these background hypotheses. We illustrate the resulting problematic, as we can apply neither likelihood tests nor NP tests to such hypotheses. That clearly decreases the rationality of testing such hypotheses and – at the same time – decreases the rationality of objectivist methods for statistical inference, as they have to rely on the prior validity of the background hypotheses without a proper method to test them.

We can of course test for the assumption of normality (or any other distribution) by, e.g., a Kolmogorov-Smirnov test; likewise, we can test for independence. In both cases, the tests are significance tests in Fisher's sense without a possibility for integrating alternative hypotheses. There may be some plausibility considerations whether we would fail to reject some other distribution if we failed to reject a normal distribution, to get some idea about the "power" of such a test. Yet, we cannot order the many distributions that may also serve as a suitable model for the variable under scrutiny so that power considerations to maximise the power in the NP sense are not available. Worse in case of testing independence against any kind of dependence, as there is no way to "order" various possibilities for describing dependence between the variables.

That means, for testing background hypotheses, we have to rely on significance tests, which are not seen as appropriate from the analytical-science perspective; in fact, such a test is often replaced by (subjective) arguments in favour of some models. Simplicity of the used model may here be one of the stronger arguments. Overall, a pure objectivist theory for statistical inference finally breaks down under the problem of testing background hypotheses. The specific test theories that have been investigated (NP, likelihood test theory, etc.) are not suitable for that purpose. In this light, we can no longer pursue Stegmüller's criterion of avoiding a subjectivation of science. Statistical inference shifts the connotation of probability towards an SJT connotation, be it open and formally checkable or be it hidden in private decisions by individual researchers or the community as a whole. The new challenge is, to make such decisions transparent and open to critique so that they can be improved to corroborate the used methods and develop a viable connotation of the concepts.

7 Conclusions

We investigated the project of an objectivist explication of probability. In the analytic theory of science, an important task is to find a suitable justification for an intuitive idea of a pre-concept. This justification is usually done by introducing a theory, in which it is possible to re-construct as many properties as possible for this pre-concept so that this theory allows a general approach for describing a wide spectrum of phenomena from reality and to provide potential solutions for real-world problems. The way, how one succeeds to build up this theory, the rationality, the reasonableness of required definitions, and the plausibility of additional criteria that are necessary to complete this theory, are finally a justification for the initial intuitive idea. This intuitive idea will also impregnate the phenomena to be described and will serve as a link between the real world and the theory, in which one may find solutions for problems.

There is an axiomatic justification for the SJT conception of probability, which may be described as a personal judgement about the degree of belief, which has been worked out by de Finetti (1937); this approach includes a complete inductive logic, i.e., it provides the statistical inference within the theory of probability, mainly by the Bayesian theorem. The axiomatic justification for the FQT conception of probability describes probability as linked to something like (idealised) relative frequencies in experiments that are repeatable under the same circumstances (including the independence requirement). Yet, this approach does not provide a theory for statistical inference within the undisputed mathematical theory of probability.

Starting point for our considerations were the approaches of Hacking and Stegmüller, who try to find an explication of probability on an objectivistic basis; be it a simple FQT conception, or the propensity conception of Popper (1959). That is, they refer to an objectivistic framework and try to find a satisfactory solution for statistical inference, which then in a loop, influences the probability concept. Both of them find arguments against other objectivist approaches for statistical inference and present favourable arguments for the likelihood test theory. Based on likelihood support and the likelihood principle they find the solution that combines direct inference (single case) and the judgement of hypotheses in the light of data (inverse inference). A solution, which they do not abandon even if they themselves find severe objections against the rationality of the likelihood test theory.

The paradox of Kerridge or the similar problem of Hacking, which we discussed in Section 5.3, reveal crucial flaws of the approach and shed doubt on the rationality of it. This would cause serious problems with the FQT conception, as an explication of probability based on FQT definitely fails. Yet, they do not leave this position, as the only solution left – to acknowledge the subjectivist conception of probability – is

unacceptable for them, as this would lead to a subjectivation of the natural sciences where an FQT probability is essential.

We have used the same paradoxes to criticise the objectivist conception per se, as we see – by a formal use of (SJT) prior probabilities – why the likelihood argument has to fail. We do not advocate switching to a closed SJT position of probability, however. We just seek for a transparent use of probability and probability models, which always have a genuine subjectivist component. A reduction of the concept of probability either to an objectivist or to a subjectivist conception would distort the concept as it has emerged in parallel due to the tension between both connotations. The debate in the foundations – led by the paradigm of science, especially the objective paradigm of physics – forbade accepting the SJT approach for probability and accepted an approach towards statistical inference with many rationality problems. It even accepted Popper's (1962/1935) view that one can only continuously test statistical hypotheses so that they would be prone to be rejected. That method would thus overall enrich the proportion of correct hypotheses by a method of corroboration. With no aim and no possibility to make any further statements about a specific hypothesis under scrutiny. The methods of statistical inference based on FQT are highly disputed up to date.

As for the didactical implications of our analysis, we regard didactic as a discipline that combines knowledge and insights from mathematics and any other science, which could support teaching and understanding; we may list here (cognitive) psychology, sociology, philosophy of science, and many others. Different from empirical investigations in how methods would be effective in actual teaching and which factors would influence it most, the hermeneutic method elaborates on a topic by gathering arguments, facts, and analogies to identify problems, to compare approaches towards the situation, or to analyse how an approach frames the situation.

An open consideration of the peculiarities of the methods around Bayes' formula and a transparent account for subjectivist and objectivist parts of the concept of probability would highlight its mutual character. Probability lives from a complementarity of both constituents, it is an entity that has an empirical counterpart and it is an entity that forms our thought about randomness so strongly that we now can understand why de Finetti (1974) stated "PROBABILITY DOES NOT EXIST. IT IS ONLY IN OUR MIND". Even though this is an extreme statement of an extreme adherent of the SJT conception, the citation should mark the one end of probability, which – so far – is less known and less well accepted in comparison to the FQT view. It is the complementarity between these diverse meanings that shapes our thought about randomness and related mathematical models.

The present paper focuses on the ways a frequentist interpretation of probability can be used to develop suitable methods for statistical inference. The discussion about the

controversy in the foundations revealed that a frequentist conception is highly prone to dispute, as a justification of an FQT meaning failed from a rational perspective when the explication of probability integrates statistical inference. This gives rise to investigating refined approaches towards teaching probability and statistics from a wider perspective on the range of meanings of probability apart from FQT. Carranza and Kuzniak (2008) provide evidence about shortcomings of educational approaches that ignore SJT aspects of probability, which lead to far-reaching misconceptions not only about the use of Bayes' formula but in the perception of probabilities at large. Anyway, we have argued that the concept of probability emerges from a mixture between APT, FQT, and SJT meanings, which are not easy to separate. The shift in connotation of probability towards a biased FQT meaning decreases the scope of probability or the quality of applications. Probability is a complementary concept, which falls apart if we reduce it to one view. Steinbring (1991) speaks of the "theoretical character of probability" in this connection.

References

[1] ARBUTHNOT, J. An argument for divine providence taken from the constant regularity observed in the birth of both sexes. Philosophical Transactions of the Royal Society, v. 27, p. 186-190, 1712.

[2] BARNETT, V. Comparative statistical inference. New York: Wiley, 1982.

[3] BATANERO, C.; BOROVCNIK, M. Statistics and probability in high school. Rotterdam: Sense Publishers, 2016.

[4] BATANERO, C.; CHERNOFF, E.; ENGEL, J.; LEE, H.; SÁNCHEZ, E. Research on teaching and learning probability. ICME-13 Topical Surveys. Cham: Springer online, 2016.

[5] BATANERO, C., HENRY, M., PARZYSZ, B. The nature of chance and probability. In: JONES, A. G. Exploring probability in school. New York: Springer, 2005. p. 15-37.

[6] BAYES, T. An essay towards solving a problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society, v. 53, p. 370-418, 1763.

[7] BELLHOUSE, D. R. De Vetula: A medieval manuscript containing probability calculations. International Statistical Review, v. 68, n. 2, p. 123-136, 2000.

[8] BERNOULLI, J. Ars conjectandi. Basel: Impensis Thurnisiorum, Fratrum, 1713.

[9] BOHR, N. The quantum postulate and the recent development of atomic theory. Nature, v. 121 n. 3050, p. 580-590, 1928.

[10] BOREL, E. Les probabilités et la vie. Paris: Presses Universitaires de France, 1943.

[11] BOROVCNIK, M. Was bedeuten statistische Aussagen. Vienna: Hölder-Pichler-Tempsky, 1984.

[12] BOROVCNIK, M. Probabilistic and statistical thinking. In Bosch, M. Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education. Barcelona: European Society for Research in Mathematics Education, 2006. p. 484-506.

[13] BOROVCNIK, M. Risk and decision making: The "logic" of probability. The Mathematics Enthusiast, v. 12. n. 1, 2 & 3, p. 113-139, 2015.

[14] BOROVCNIK, M.; KAPADIA, R. A historical and philosophical perspective on probability. In: Chernoff, E. J.; Sriraman, B. Probabilistic thinking: Presenting plural perspectives. New York: Springer, 2014. p. 7-34.

[15] BOX, G. E. P.; DRAPER, N. R. Empirical model-building and response surfaces. New York: Wiley, 1987.

[16] CARRANZA, P.; KUZNIAK, A. Duality of probability and statistics teaching in French education. In: BATANERO, C.; BURRILL, G.; READING, C.; ROSSMAN, A. Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Monterrey: ICMI and IASE, 2008.

[17] ÇINLAR, E. Probability and stochastics. Berlin, New York: Springer, 2011.

[18] COLQUHOUN, D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, v. 1, p. 1-16, 2014.

[19] DAVID, F. N. Games, gods and gambling. London: Griffin, 1962.

[20] FINETTI, B. de. La prévision: ses lois logiques, ses sources subjectives. Annales Institut Henri Poincaré, v. 7, p. 1-68, 1937; Foresight: Its logical laws, its subjective sources. In: KOTZ, S.; JOHNSON, N. L. Breakthroughs in statistics. Vol. I. New York: Springer, 1992. p. 134-174.

[21] FINETTI, B. de. Theory of probability. New York: Wiley, 1974.

[22] FISHER, R. A. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.

[23] FISHER, R. A. The design of experiments. Edinburgh: Oliver and Boyd, 1935.

[24] GIGERENZER, G. Die Evolution des statistischen Denkens. Stochastik in der Schule, v. 24, n. 2, p. 2-13, 2004.

[25] GOOD, I. J. Good thinking. The foundations of probability and its applications. Mineola, NY: Dover Publications, 1983.

[26] GORARD, S.; WHITE, P. Still against inferential statistics: Rejoinder to Nicholson and Ridgway. Statistics Education Research Journal, v. 16, n. 1, p. 70-75, 2017.

[27] GUTTMAN, L. The illogic of statistical inference for cumulative science. Applied Stochastic Models and Data Analysis, v. 1, p. 3-10, 1985.

[28] HACKING, I. The logic of statistical inference. Cambridge: Cambridge Univ. Press, 1965.

[29] HACKING, I. The emergence of probability. Cambridge: Cambridge Univ. Press, 1975.

[30] HILBERT, D. Mathematische Probleme. Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, p. 253–297, 1900.

[31] HUNTER, J. E. Needed: A ban on the significance test. Psychological Science, v. 8, n. 1, p. 3-7, 1997.

[32] JEFFREYS, H. Theory of probability. 2nd Ed. Oxford: Clarendon, 1948.

[33] KAPADIA, R.; BOROVCNIK M. Chance encounters. Dordrecht: Kluwer, 1991.

[34] KOLMOGOROV, A. N. Grundbegriffe der Wahrscheinlichkeitsrechnung. Zentralblatt für Mathematik, v. 2, n. 3, 1933. Reprinted: Berlin: Springer, 1977.

[35] KOLMOGOROV, A. N. Foundations of the theory of probability. New York: Chelsea, 1956.

[36] LANE, D. A. Fisher, Jeffreys, and the nature of probability. In: FIENBERG, S. E.; HINKLEY, D. V. (Eds.). R.A. Fisher: An appreciation. New York: Springer, 1980, p. 140-160.

[37] LAPLACE, P. S. de Essai philosophique sur les probabilités. Journal de l'École Polytechnique, v. VII/VIII, p. 140-172, 1812.

[38] MISES, R. v. Grundlagen der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift, v. 5, p. 52-99, 1919.

[39] MULAIK, S. A.; RAJU, N. S.; HARSHMAN, R. A. There is a time and a place for significance testing. In: HARLOW, L. L.; MULAIK, S. A.; STEIGER, J. H. What if there were no significance tests? Brighton: Psychology Press, 1997. p. 61-106.

[40] NEYMAN, J. Outline of a theory of statistical estimation based on the classical theory of probability. Transactions of the Royal Statistical Society, v. 97, p. 558-625, 1937.

[41] NEYMAN, J.; PEARSON, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I and II. Biometrika, v. 20A, p. 175-240; p. 263-294, 1928.

[42] NICKERSON, R. S. Null hypothesis significance testing: A review of an old and continuing controversy. Psychological Methods, v. 5, n. 2, p. 241-301, 2000.

[43] OTTE, M. Ways of knowing and modes of presentation. In: CIEAEM. Moyens et medias dans l'enseignement des mathematiques. Orleans: Université, 1984. p.41-69.

[44] PEARSON, E. S. The selected papers of E. S. Pearson. Berkeley: University of California Press, 1966.

[45] POPPER, K. R. The propensity interpretation of probability. British Journal of the Philosophy of Science, v. 10, p. 25-42, 1959.

[46] POPPER, K. R. Logic of scientific discovery (English version of Logik der Forschung). London: Routledge, 1962/1935.

[47] SAVAGE, L. J. The foundation of statistical inference. London: Methuen, 1962.

[48] SAVAGE, L. J. On rereading R. A. Fisher. The Annals of Statistics, v. 4, p. 441-500, 1976.

[49] SEIDENFELD, T. Philosophical problems of statistical inference – Learning from R. A. Fisher. Dordrecht: D. Reidel, 1979.

[50] STEGMÜLLER, W. Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie, Vol. 4, 1st part: Personelle Wahrscheinlichkeit und Rationale Entscheidung, 2nd part: Personelle und statistische Wahrscheinlichkeit. Berlin-New York: Springer, 1973.

[51] STEINBRING, H. Zur Entwicklung des Wahrscheinlichkeitsbegriffs – Das Anwendungsproblem in der Wahrscheinlichkeitstheorie aus didaktischer Sicht. Bielefeld: IDM, 1980.

[52] STEINBRING, H. The theoretical nature of probability in the classroom. In Kapadia, R.; Borovcnik, M. Chance encounters. Dordrecht: Kluwer, 1991. p. 135-168.

[53] VENN, J. The logic of chance. Reprinted. New York: Chelsea, 1866/1962.

Submitted on 20 November 2020.
Accepted on 21 August 2021.