The philosopher Popper emphasized that ‘induction is a myth’. This philosophically oriented, or one could say rational, point of view seems to conflict the more empirically oriented opinion, shared by many research workers, that ‘induction is a must’. The saying that inductive inferences cannot be ‘logically valid’ refers to a narrow interpretation of logical validity. Of course, every research worker who makes inductive inferences realizes that he falls short with respect to reality. The question of interest is, however, whether or not the inferences are reliable from the viewpoint of truth approximation. Reichenbach [72] summarized this difference of perspective by writing ‘induction is not logic, it is a habit’. The discussion of the inferential habits, whether they are good or bad, goes beyond logic and involves facts (data) as well as fictions (models and principles). Different choices of models or principles will inevitably lead to different inferences. In the theory of statistical inference, the validity of inferences is deduced from the fact that they are constructed by a method that is, in some sense, optimal. To define optimality one has to formulate the underlying principles.

The Dutch statistician Van Dantzig made clear that the basis of inferential principles should be the pursuit of cultural values which he tried to indicate with the Platonic words Truth, Beauty and Goodness. As these ideals are unattainable for any statistical theory, one can only try to approximate them as good as possible. To this end, one could emphasize different aspects of these cultural values; a strong emphasis on Beauty will lead to theory that is dominated by mathematical coherence, whereas a strong emphasis on Truth will lead to the awareness that it are the data which should speak and that this should be allowed by the context which should not be too rigid. The art
consists, of course, in trying to balance these two extremes. It is in this respect that some form of Goodness will be pursued.

1.1 The statistical controversy

The history of statistical inference is marked by controversies about its fundamental principles. Wilkinson [95] wrote about these controversies:

‘In considering why the controversy has continued for so long, one notices first a peculiar difficulty in invalidating a possibly erroneous theory of uncertain inference. Scientific theories about Nature ultimately stand or fall in the basis of clear observational proof (or otherwise) of predicted properties of Nature. In case of uncertain inference, however, the very uncertainty of uncertain predictions renders the question of their proof or disproof almost meaningless. Invalidation of an inference theory therefore depends on the discovery of extreme examples from which the derived statements of uncertainty are clearly incorrect from an intuitive or empirical point of view.’

The first part of this statement clearly shows the need for rules to evaluate uncertain inferences. Hence, providing tools for evaluation is one of the first tasks if one wants to discuss the various approaches to statistical inference. The second part of this statement contains a warning against being too dogmatic; an axiomatic approach to statistical inference will often lead to paradoxical situations.

Historically, one can distinguish roughly four broad approaches to statistical inference. This division into different schools is mainly on the basis of the way their principles approximate the cultural values, in the sense that they give different weights to them. A description of the main features of the four mainstream statistical schools and the differences between them can be found in, e.g. Cox [22], [23]. This description is literally the following:

(i) ‘The rather ill-defined but rich set of ideas that is conveniently called Fisherian, after the geneticist and statistician R.A. Fisher. Fisher explicitly emphasized the need for a variety of approaches for different problems; he was dismissive of axiomatic arguments.

(ii) An approach strongly based on a frequency theory of probability, initially developed to explain Fisher’s ideas more concretely, and emphasizing operational
1.1. The statistical controversy

concepts. This Neyman–Pearson theory, is that presented most commonly in textbooks and courses.

(iii) An approach in which probability represents a rational degree of belief, i.e. $P(A|'data', H)$ measures the degree of belief a reasonable person would have in $A$, given the data under analysis and additional information represented by $H$. This goes back to Laplace and his predecessors; in its modern form it is associated with the philosopher R. Carnap and especially with the geophysicist and applied mathematician H. Jeffreys.

(iv) An approach in which probability represents your degree of belief constrained by a requirement of self-consistency. This is associated with F.P. Ramsey, I.J. Good, B. de Finetti and L.J. Savage. There is no assumption that different people faced by the same evidence have the same probability.

In the first two approaches the procedures are justified by their performance under hypothetical repetition of the experiment, i.e. frequency properties. The differences between the Fisherian and the Neyman–Pearson approaches are minor, and are essentially the following.

(i) In the Fisherian approach emphasis is placed on the simple test of significance, on the likelihood function and principles as sufficiency and interval estimation through fiducial distributions.

(ii) The Neyman–Pearson approach emphasizes operational requirements such as power and other explicit indicators of sensitivity, and uses procedure oriented terminology such as acceptance and rejection of hypotheses.

(iii) There is a difference in the attitude to what philosophers call the problem of the unique case.

Jeffreys's approach to the issue of inference has the same target as Fisher's: what can we reasonably learn about a parameter in the light of the model for the data? In contrast to Fisher, Jeffreys argues that a different notion of probability, that of a reasonable degree of belief, is needed to achieve this. He comes to the conclusion that these probabilities should satisfy the axioms of probability theory, and therefore have to be computed by Bayes's rule. The prior distribution is taken, in accordance with Laplace, to be very dispersed, representing lack of knowledge. In many situations Jeffreys's approach provides the same answers as the Fisherian approach. Jeffreys's and the personalistic approach are often referred to as Bayesian approaches. Although they are formally the same, there are some philosophical differences, namely the following.

(i) The personalistic degree of belief measures, in contrast to the reasonable degree of belief, describes how strongly you believe in $A$ in the light of the model
for the data. As a consequence, the choice of prior is often different.

(ii) The probabilistic coherence of the personalistic probabilities is justified by thinking in terms of betting strategies.'

To avoid dogmatism that can lead to paradoxic situations, an eclectic approach that combines the various theories of statistical inference is needed. Such an approach, however, has the disadvantage that it leaves a substantial methodological uncertainty. If there are large amounts of data available, then the different approaches will usually lead to the same numerical answers, though their interpretation will still depend on the paradigm by which they are generated. For small samples, however, the numerical answers can differ substantially. In the latter case guidelines are needed to reach a compromise. Sometimes the methodological uncertainty is so overwhelming that one serious possibility, that is often overlooked, is to return the problem to the ‘owner’ with the remark that it cannot be settled by the profession.

To reach a compromise between the various approaches for making inference, it is necessary to be able to pinpoint the similarities and differences between the various approaches. It is the aim of the theory of *distributional inference* to create a mathematical framework in which the various approaches have their place, and to provide tools for making comparative analyses. The term ‘distributional inference’ refers to the required form of inference, namely, that of a probability distribution on the space of the theoretically possible values of the unknown of interest. This distributional form of inference corresponds to that pursued by Fisherian and Bayesian analyses of the problem. The link between the distributional form of inference and the confidence regions arising from the Neyman–Pearson approach is provided by thinking of a probability distribution on the space of the theoretically possible values in terms of a coherent family of confidence intervals of all levels.

The theory of distributional inference has a close resemblance with the theory of point estimation. First, prescriptions are given for evaluating inferences, and subsequently for evaluating procedures for making inferences by using hypothetical repetition of the experiment. Next, procedures are compared on the basis of different performance characteristics. Finally, a qualitative discussion is given to reach a compromise.
1.2 Statistical hypotheses

It is a fundamental problem in statistical inference to discuss the truth or falsity of statistical hypotheses. With the work of Thomas Bayes [6](1702–1761), published posthumously by Price in 1763, this problem began to attract serious attention. Bayes was one of the first who attempted to apply the notion of probability in order to relate causes to their effects. In 1774 Laplace formulated a simplified version of what is presently known as Bayes's Theorem: if \( A_1, \ldots, A_k \) are the causes considered possible in relation to the event \( X \), then the probability of \( A_i \) given that the event \( X \) occurred is determined by

\[
P(A_i|X) = \frac{P(X|A_i)}{\sum_{j=1}^{k} P(X|A_j)}.
\]

This is precisely Bayes's Theorem in the special case of equal prior probabilities. In this fashion, Laplace, De Morgan, Boole, and many others calculated posterior probabilities of two mutually exclusive statements (hypotheses). However, the question as to whether it is justified to assign such prior probabilities to such causes remained, see e.g. Stigler [86].

Satisfactory alternatives became available when Helmert [47], and especially Pearson [69], started to use the probability of the tail of the null distribution as degree of belief in the hypothesis. Then Fisher made an important contribution by deriving many exact null distributions. This led to the situation where the computation of \( P \)-values was preferred to the computation of posterior probabilities. The choice of the underlying test statistic, however, remained a matter of intuition. Neyman and E.S. Pearson replaced this intuition by their likelihood ratio principle.

To support this likelihood ratio principle, Neyman and Pearson shifted the attention from the computing of significance probabilities to the making of accept–reject decisions. They focused on constructing best critical regions under the restriction that the probability of an error of the first kind does not exceed some nominal level. These ideas were the basis of the procedure-oriented or behavioral approach which was given form by Wald. The mathematical simplicity of the accept–reject formulation made the Neyman–Pearson approach attractive for mathematical–statistical investigation. For general scientific purposes, however, such accept–reject decisions will be too crude: a more precise indication is needed to express the extent at which the data corroborates the hypothesis. In his standard work about the Neyman–Pearson theory, Lehmann [58] writes (a slight notational modification has been made):

\[
P(A_i|X) = \frac{P(X|A_i)}{\sum_{j=1}^{k} P(X|A_j)}.
\]
In applications there is usually available a nested family of rejection regions, corresponding to different significance levels. It is then good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also the smallest significance level, the significance probability $\alpha = \alpha_p(x)$ or P-value, at which the hypothesis would be rejected for the given observation.

In this respect the P-value occurs as a byproduct of the Neyman–Pearson theory. It is used as a measure of the evidence carried by the data in favor of the hypothesis.

In the seventies and eighties the neo–Bayesian school emerged, firstly under the leadership of L.J. Savage. They criticized the Neyman–Pearson–Wald approach and concentrated on developing a theory of personalistic probabilities. This resulted in a revival of posterior probabilities as degrees of belief in the hypotheses. At the mid–eighties, authors like Berger [10] and Kiefer [53] reconsidered aspects of the Neyman–Pearson theory and provided other alternatives to the P-value. These were based on posterior probabilities, but were not posterior probabilities themselves. Dissatisfaction with the underlying theory was motivation for the development of a theory of Q-values (Schaafsma [79]). Such Q-values are estimators for the indicator function of the hypothesis, and are based on squared error (Quadratic) loss.

The idea that a degree of belief in the hypothesis is an estimate of the true value of the corresponding indicator function provides a mathematical framework that encompasses the proposals of the various schools. Within this framework it is possible to evaluate and compare the different approaches. In the second chapter of this thesis, this idea will be explored for problems of discussing the truth or falsity of point null hypotheses, in the third chapter the theory will be worked out for problems concerning general hypotheses. Special attention will be paid to the place of the P-value within this framework, because this provides a link to the theory of fiducial inference.

1.3 Fiducial inference

In his paper ‘Inverse Probability’, Fisher [37] introduced the fiducial argument as a method to generate distributional inferences without specifying a prior. He illustrated his argument by making distributional inferences about the correlation coefficient of a bivariate normal random sample. He may have chosen this example because of a long standing conflict with Karl Pearson. At the end of Chapter 5 this problem will be reexamined. For a while, Fisher thought
that the fiducial argument provides instances of an inductive logic. He derived fiducial inferences for a number of problems (see e.g. Savage [76]), without being sufficiently clear about the underlying principles. Though the fiducial argument has been subject of a lot of criticism, papers by Wilkinson [95], Pedersen [70], Buehler [18], Dawid–Stone [25], Wallace [92], Seidenfeld [82], Zabell [96], Barnard [4], Kroese et al. [55], [56], and Fraser [41] indicate that the interest in it has not vanished. Zabell concluded his review as follows:

‘Fiducial inference (...) was in essence an attempt to (...) ‘make the Bayesian omelette without breaking Bayesian eggs’ (...) Fisher’s attempt to steer a path between the Scylla of unconditional, behaviourist methods which disavow any attempt at ‘inference’ and the Charybdis of subjectivism in science was founded on important concerns, and his personal failure to arrive at a satisfactory solution to the problem means only that the problem remains unsolved, not that it does not exist.’

Bartlett indicated that fiducial methods may have certain deficiencies but Fisher did not accept the criticism. This caused an alienation between Fisher and Neyman who insisted upon the requirement that fiducial limits should correspond to exact confidence statements. Barnard wrote about the Fisher–Neyman conflict as follows:

‘any hopes that further discussion (...) might eventually result in better mutual understanding between these two great men were dashed by the outbreak of war in 1939, which scattered the leading personalities far and wide and set them to work on a variety of more immediate tasks.’

The unresolved issues have had the effect that Neyman and his followers focused on confidence intervals, tolerance intervals, etc., and deprived themselves from the elegance which the distributional form of inference provides. The situation is still much like it was when Wallace wrote:

‘Of all R.A. Fisher’s creations, fiducial inference was perhaps the most ambitious, yet least accepted. Ignored by a large part of the statistics community satisfied with the mathematically simpler confidence approach, and rejected as logically imperfect and inconsistent in general by those who recognized the strength of the fiducial objectives, the fiducial argument continues under active and sympathetic study today only in a few islands of the statistical world’.
Barnard refers to the need for a ‘valid mode of reasoning of wide applicability which allows us to make probability statements other than personal ones concerning some of the parameters with which the data analysts have to deal’. A related suggestion was made by James in his comment on Wilkinson [95]:

‘I think that R.A.Fisher’s postulate of ignorance for fiducial theory should be replaced by a positive requirement that the statistical analysis is appropriate to and meets the issues at stake.’

In the fourth chapter of this thesis the original formulation of the fiducial argument will be investigated, and the modern extensions of fiducial theory will be reviewed. Next, in the fifth chapter, fiducial inference will be placed in the more general context of making distributional inferences. This makes it possible to compare it with other paradigms for generating distributional inferences, for example the Bayesian one.

1.4 The plan

The outline of this thesis is as follows. In the second chapter the problem of discussing the truth or falsity of point null hypotheses will be considered. The P-value, regarded as an estimator of the indicator function of the hypothesis, will be compared to other estimators, e.g. posterior probabilities, on the basis of squared error loss. An important ingredient in this comparison is the restriction to so called ‘unbiased procedures’. Two related types of unbiasedness will be considered; weak and strong unbiasedness. Both concepts of unbiasedness are more closely related to unbiasedness concepts from the Neyman–Pearson theory of testing statistical hypotheses than to the concept of mean unbiasedness, which is commonly used in the theory of point estimation.

In the third chapter, the ideas developed in the second chapter will be formalized, and the scope will be extended to general testing problems. Relations between properties of risk functions and weak and strong unbiasedness are investigated. It will be shown that, for some problems, the P-value has some intrinsic optimality property. Special emphasis will be given to one-sided testing problems, because they can be used as building blocks for making distributional inferences about a real-valued parameter.

Next, in the fourth chapter, the fiducial argument will be introduced and given a modern interpretation in terms of structured models. Within this context of structured models, some results about fiducial inference will be reviewed. For example, it will be shown that, in contrast to Bayesian inferences, fiducial
inferences are not probabilistically coherent. Moreover, the connection between
fiducial inference and Bayesian inference is investigated in the case that the
model is invariant under the action of an algebraic group.
In the fifth chapter, the theory of chapter three is extended in the context
of making distributional inferences about real–valued parameters. An interpre-
tation of the fiducial argument is given providing distributional inferences by
combining P–values corresponding to one–sided testing problems. This inter-
pretation is related to that of Neyman who constructed confidence regions as
intervals of nonrejected null hypotheses. Decision–theoretic properties of fidu-
cial inferences can be derived from the results for P–values in Chapter 3.
In the last chapter, the theory will be applied to the problem of making
inferences about the quantiles of an unknown probability distribution. This
will shed new light on existing quantile estimators and provide a possibility to
derive a nonparametric density estimator which is not based on one of the usual
methods.