Inductive Logic and Statistics

Romeijn, J. -W.

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2009

Inductive Logic and Statistics

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen
j.w.romeijn@rug.nl

Abstract

This chapter[1] concerns inductive logic in relation to mathematical statistics. I start by introducing a general notion of probabilistic inductive inference. Then I introduce Carnapian inductive logic, and I show that it can be related to Bayesian statistical inference via de Finetti's representation theorem. This in turn suggests how Carnapian inductive logic can be extended to include inferences over statistical hypotheses. With this extension inductive logic becomes more easily applicable to statistics. I consider two classical statistical procedures, maximum likelihood estimation and Neyman-Pearson hypothesis testing, and I discuss how they can be accommodated in an inductive logic with hypotheses.

1 From inductive logic to statistics

There are strong parallels between statistics and inductive logic. An inductive logic is a system of inference that describes the relation between data statements, and statements that extend beyond the data, such as predictions over future data, and general conclusions on all possible data. Statistics, on the other hand, is a mathematical discipline that describes procedures for deriving results about a population from sample data. These results include decisions on rejecting or accepting a hypothesis about the population, the

[1]This chapter was written in parallel with a chapter for the Handbook for the Philosophy of Science: Philosophy of Statistics, edited by Bandyopadhyay and Forster (2009). The two chapters show considerable overlap. The present chapter considers how inductive logic can be developed to encompass the statistical procedures. The other chapter approaches the same material from the other end, and aims at a reconstruction of statistical procedures in terms of inductive logics.

determination of probability assignments over such hypotheses, predictions on future samples, and so on. Both inductive logic and statistics are thus calculi for getting from the given data to statements or results that transcend the data.

Despite this fact, inductive logic and statistics have evolved more or less separately. This is partly because there are objections to viewing classical statistics as inferential. Another reason may be that inductive logic has been dominated by the Carnapian programme, and that statisticians have perhaps not recognised inductive logic as a discipline that is much like their own. Statistical hypotheses and models do not appear in the latter, but they are the start and finish of most statistical procedures. Against this, I aim to show that Carnapian inductive logic can be developed to encompass inference over statistical hypotheses, and that the resulting inductive logic can, at least partly, capture statistical procedures.

In doing so, I hope to bring the philosophical discipline of inductive logic and the mathematical discipline of statistics closer together. I believe both disciplines can benefit from such a rapprochement. First, framing statistical procedures as inferences in an inductive logic may help to clarify the presuppositions and foundations of these procedures. Second, by relating statistics to inductive logic, insights from inductive logic may be used to enrich and improve statistics. And finally, showing the parallels between inductive logic and statistics may show the relevance, also to inductive logicians themselves, of their discipline to the sciences, and thereby direct further research in this field.

The reader may wonder where in this chapter she can read about the history of inductive logic in relation to the historical development of statistics. Admittedly, positions and theories from both disciplines are here discussed from a systematic viewpoint, and not so much as historical entities. I aim to provide a unified picture of inductive inference to which both inductive logic and statistics, past or present, can be related. At the heart of this picture lies the notion of statistical hypothesis. I think the fact that inductive logic and statistics have had comparatively little common past can be traced back to the absence of this notion from inductive logic. In turn, this absence can be traced back to the roots of inductive logic in logical empiricism. In that derived sense, the exposition of this chapter is related to the history of inductive logic.

The plan of the chapter is as follows. I start by describing induction and observations in formal terms. Then I introduce a general notion of probabilistic inductive inference over these observations. Following that I present Carnapian inductive logic, and I show that it can be related to Bayesian statistical inference via de Finetti's representation theorem. This in turn suggests how Carnapian inductive logic can be extended to include inferences over statistical hypotheses. Finally, I consider two classical statistical procedures, maximum likelihood estimation and Neyman-Pearson hypothesis testing, and I discuss how they can be accommodated in this extended inductive logic.

Given the nature of the chapter, the discussion of statistical procedures is relatively short. Many statistical procedures are not dealt with. Similarly, I cannot discuss in detail the many inductive logics devised within Carnapian inductive logic. For the latter, the reader may consult chapters 9 and 10 in this volume, and the further references contained therein. For the latter, I refer to a recent volume on the philosophy of statistics, edited by Bandyopadhyay and Forster (2009).

2 Observational data

As indicated, inductive inference starts from statements on data, and ends in statements that extend beyond the data. An example of an inductive inference is that, from the statement that up until now all observed pears were green, we conclude that the next few pears will be green as well. Another example is that from the green pears we have seen we conclude that all pears are green, period. The key characteristic is that the conclusion says more than what is classically entailed by the premises.

Let me straighten the inferences out a bit. First, I restrict attention to statements on empirical facts, thus leaving aside such statements as that pears are healthy, or that God made them. Second, I focus on the results of observations of particular kinds of empirical fact. For example, the empirical fact at issue is the colour of pears, and the results of the observations are therefore colours of individual pears. There can in principle be an infinity of such observation results, but what I call data is always a finite sequence of them. Third, the result of an observation is always one from a designated partition of properties, usually finite but always countable. In the pear case, it may be {red, green, yellow}. I leave aside observations that cannot be classified in terms of a mutually exclusive set of properties.

I now make these ideas on what counts as data a bit more formal. The concept I want to get across is that of a sample space, in which single observations and sequences of observations can be represented as sets, sometimes called events. After introducing the observations in terms of a language, I define sample space. All the probabilities in this chapter will be defined over that space because, strictly speaking, probability is axiomatized as a measure function over sets. However, the expressions may be taken as sentences just as well.

We denote the observation of individual i by Q_i. This is a propositional variable, and we denote assignments or valuations of this variable by q_i^k, which represents the sentence that the result of observing individual i is the property k. A sequence of such results of length t, starting at 1, is denoted with the propositional variable S_t, with the assignment $s^{k_1 \ldots k_t}$, often abbreviated as s_t. In order to simplify notation, I denote properties with natural numbers, so $k \in K = \{0, 1, \ldots, n-1\}$. For example, if the observations are the aforementioned colours of pears, then $n = 3$. I write red as 0, green as 1, and yellow as 2, so that s^{012} means that the first three pairs were red, green, and yellow respectively. Note further that there are logical relations among the sentences, like $s^{012} \rightarrow q_2^1$. Together, the expressions s_t and q_i^k form the observation language.

It will be convenient to employ a set-theoretical representation of the observations, a so-called sample space, otherwise known as an observation algebra. To this aim, consider the set of all infinitely long sequences K^Ω, that is, all sequences like $012002010211112\ldots$, each encoding the observations of infinitely many pears. Denote such sequences with ω, and write $\omega(i)$ for the i-th element in the sequence ω. Every sentence q_i^k can then be associated with a particular set of such sequences, namely the set of ω whose i-th element is k:

$$q_i^k = \{\omega \in K^\Omega : \omega(i) = k\}.$$

Clearly, we can build up all finite sequences of results $s^{k_1 \ldots k_t}$ as intersections of such sets:

$$s^{k_1 \ldots k_t} = \bigcap_{i=1}^{t} q_i^{k_i}.$$

Note that entailments in the language now come out as set inclusions: we have $s^{012} \subset q_2^1$. Instead of using a language with sentences q_i^k and logical relations among such sentences, I will in the following use a so-called algebra $\mathcal{Q}$, built up by the sets q_i^k and their conjunctions and intersections.

Finally, I want to emphasise that the notion of a sample space introduced here is really quite general. It excludes a continuum of individuals and a continuum of properties, but apart from that, any data recording that involves individuals and that ranges over a set of properties can serve as input. For example, instead of pears having colours we may think of subjects having test scores. Or of companies having certain stock prices. The sample space used in this chapter follows the basic structure of most applications in statistics, and of almost all applications in inductive logic.

3 Inductive inference

Now that I have made the notion of data more precise, let me turn to inductive inference. Consider the case in which I have observed three red pears: s^{000}. What can I conclude about the next pear? Or about pears in general? From the structure of the data itself, it seems that we can conclude depressingly little. We might say that the next pear is red, q_4^0. But as it stands, each of the sets $s^{000k} = s^{000} \cap q_4^k$, for $k = 0, 1, 2$, is a member of the sample space. The event of observing three red pears is consistent with any colour for the next pear. Purely on the basis of the classical relations among observations, as captured by the sample space, we cannot draw any inductive conclusion.

Perhaps we can say that given three green pears, the next pear being red is more probable? This is where we enter the domain of probabilistic inductive logic. We can describe the complete population of pears by a probability function over the observational facts,

$$P : \mathcal{Q} \mapsto [0, 1].$$

Every possible pear q_{t+1}^k, and also every sequence of such pears $s^{k_1 \dots k_t}$, receives a distinct probability. The probability of the next pear being of a certain colour, conditional on a given sequence, is expressed as $P(q_{t+1}^k | s^{k_1 \dots k_t})$. Similarly, we may wonder about the probability that all pears are green, which is again determined by the probability assignment, in this case $P(\{\forall i :$

$q_i^1\}$). All such inductive inferences are completely determined by the full probability function P.

The central question of any inductive inference or procedure is therefore how to determine the function P, relative to the data that we already have. What must the probability of the next observation be, given a sequence of observations gone before? And what is the right distribution over all observations, given the sequence? In the framework of this chapter, both statistics and inductive logic aim to provide an answer to these questions, but they do so in different ways.

It will be convenient to keep in mind a particular understanding of probability assignments P over the sample space, or observation algebra, $\mathcal{Q}$. Recall that in classical two-valued logic, a model of the premises is a complete truth valuation over the language, subject to the rules of logic. Because of the correspondence between language and algebra, the model is also a complete function over the algebra, taking the values $\{0, 1\}$. By analogy, we may consider a probability function over an observation algebra, which takes the values in the interval $[0, 1]$ and which is subject to the axioms of probability, as a model too.

In the following I will use probability functions over sample space as models, that is, as the building blocks of a formal semantics. But we must be careful with the terminology here, because in statistics, models often refer to sets of statistical hypotheses. In the following, I will therefore refer to complete probability functions over the algebra as *statistical hypotheses*. A hypothesis is denoted h, the associated probability function is P_h. In statistics, these probability functions are also often referred to as *distributions* over a population.

Probabilistic inductive logics use probability functions over sample space for the purpose of inductive inference. But there are widely different ways of understanding the inductive inferential step. The most straightforward of these, and the one that is closest to statistical practice, is to map each sample s_t onto a hypothesis h, or otherwise onto a set of such hypotheses. The inferential step then runs from the data s_t and a set of statistical hypotheses, each associated with a probability function P_h, towards a more restricted set, or even to a single $h^\star$ and $P_{h^\star}$. The resulting inductive logic is *ampliative*, because the restriction on the set of probability functions that is effected by the data, i.e. the conclusion, is often stronger than what follows from the data and the initial set of probability functions, i.e. the premises, by deduction.

We can also make the inferential step precise by analogy to a more classical, *non-ampliative* notion of entailment. As will become apparent, this kind of inferential step is more naturally associated with what is traditionally called inductive logic. It is also associated with a basic kind of probabilistic logic, as elaborated in Hailperin (1996) and more recently in Haenni et al. (2008), especially section 2. Finally, this kind of inference is strongly related to Bayesian logic, as advocated by Howson (2003). It is the kind of inductive logic favored in this chapter.

Recall that an argument is classically valid if and only if the set of models satisfying the premises are contained in the set of models satisfying the conclusion. The same idea of classical entailment may now be applied to the probabilistic models over sample space. In that case, the inferential step is from one set of probability assignments, characterised by a number of restrictions associated with premises, towards another set of probability assignments, characterised by a different restriction that is associated with a conclusion. The inference is then valid if the former is contained in the latter. In such a valid inferential step, the conclusion does not amplify the premises.

As an example, say that we fix $P(q_1^0) = 1/2$ and $P(q_1^1) = 1/3$. Both these probability assignments can be taken as premises in a logical argument, and the models of these premises are simply all probability functions P over $\mathcal{Q}$ for which these two valuations hold. By the axioms of probability, we can derive that any such function P will also satisfy $P(q_1^2) = 1/6$. On its own, the latter expression amounts to a set of probability functions over the sample space $\mathcal{Q}$ in which the probability functions that satisfy both premises are included. In other words, the latter assignment is classically entailed by the two premises.

Along exactly the same lines, we may derive a probability assignment for a statistical hypothesis h conditional on the data s_t, written as $P(h|s_t)$, from the input probabilities $P(h)$, $P(s_t)$, and $P(s_t|h)$, using the theorem of Bayes. The classical understanding of entailment may thus be used to reason inductively, namely towards statistical hypotheses that themselves determine a probability assignment over data. In this chapter I give many examples of such Bayesian inferences.

In the following the focus will be on non-ampliative inductive logic, because Carnapian inductive logic is most easily related to non-ampliative logic. Therefore, viewing statistical procedures in this perspective makes the latter more amenable to inductive logical analysis. I do not want to claim that I thereby lay bare the real nature of the statistical procedures. Rather, I hope to show that investigating statistics along these specific logical lines clarifies and enriches statistics. Furthermore, I hope to stimulate research in inductive logic that is directed at problems in statistics.

4 Carnapian logics

With the notions of observation and induction in place, I can present the logic of induction developed by Carnap (1950, 1952). Historically, Carnapian inductive logic can lay most claim to the title of inductive logic proper. It was the first systematic study into probabilistic predictions on the basis of data.

The central concept in Carnapian inductive logic is logical probability. Recall that the sample space $\mathcal{Q}$, also called the observation algebra, corresponds to an observation language, comprising of sentences such as "the second pear is green", or in formal terms, q_2^1. The original idea of Carnap was to derive a probability assignment over the language on the basis of symmetries within the language. In the example, we have three mutually exclusive properties for each pear, and in the absence of any further knowledge, there is no reason to think of any of these properties as special or as more, or less, appropriate than the other two. The symmetry inherent to the language suggests that each of the sentences q_i^k for $k = 0, 1, 2$ should get equal probability:

$$P(q_i^0) = P(q_i^1) = P(q_i^2) = \frac{1}{3}.$$

The idea of logical probability is to fix a unique probability function over the observation language, or otherwise a strongly restricted set of such functions, on the basis of such symmetries.

Next to symmetries, the set of probability functions can also be restricted by certain predictive properties. As an example, we may feel that yellow pears are more akin to green pears, so that finding a yellow pear decreases the probability for red pears considerably, while it decreases the probability for green pears much less dramatically. That is,

$$\frac{P(q^1_{t+1}|s_{t-1} \cap q^2_t)}{P(q^0_{t+1}|s_{t-1} \cap q^2_t)} > \frac{P(q^1_{t+1}|s_{t-1})}{P(q^0_{t+1}|s_{t-1})}.$$

How such relations among properties may play a part in determining the probability assignment P is described in the literature on analogy reasoning. See Festa (2006); Maher (2000); Romeijn (2006). Interesting recent work on relations between predictive properties can also be found in Paris and Waterhouse (2008).

Any Carnapian inductive logic is thus defined by a number of symmetry principles and predictive properties, determining a probability function, or otherwise a set of such functions. One very well-known inductive logic, discussed at length in Carnap (1952), employs a probability assignment characterised by the following symmetries,

$$\begin{aligned} P(q^k_i) &= P(q^{k'}_i), & (1) \\ P(s^{k_1 \ldots k_i \ldots k_t}) &= P(s^{k_i \ldots k_1 \ldots k_t}), & (2) \end{aligned}$$

for all values of i, t, k, and k', and for all values k_i with $1 \leq i \leq t$. These two symmetries together determine a set of probability assignments P, for which we can derive the following consequences:

$$P(q^k_{t+1}|s_t) = \frac{t_k + \lambda/n}{t + \lambda}, \quad (3)$$

where n is the number of values for k, and t_k is the number of earlier instances q^k_i in the sequence s_t. The parameter $\lambda > 0$ can be chosen at will. Predictive probability assignments of this form are called Carnapian λ-rules.

The probability distributions of Equation (3) has some striking features. Most importantly, for any of the probability functions P satisfying the aforementioned symmetries, we have that

$$P(q^k_{t+1}|s_{t-1} \cap q^k_t) > P(q^k_{t+1}|s_{t-1}). \quad (4)$$

This predictive property is called instantial relevance: the occurrence of q^k_t increases the probability for q^k_{t+1}. It was a huge success for Carnap that this typically inductive effect is derivable from the symmetries alone. By providing an independent justification for these symmetries, Carnap effectively provided a justification for induction, thereby answering the age-old challenge of Hume.

Note that the outlook of Carnapian logic is very different from the outlook of the inductive logics discussed in Section 3. Any such logic starts with a set of probability functions, or hypotheses, over a sample space and then imposes a further restriction on this set, or derives consequences from it, on the basis of the data. By contrast, Carnapian logic starts with a sample space and a number of symmetry principles and predictive properties, that together fix a set of probability functions over the sample space. Just like the truth tables restrict the possible truth valuations, so do these principles restrict the logical probability functions. In sum, whereas the inductive logics above take the probability functions as given, Carnap derives them from principles that have a logical status.

If we ignore the notion of logical probability and concentrate on the inferential step, then Carnapian inductive logics fit best in the template for non-ampliative inductive logic. As said, we fix a set of probability assignments over the sample space by means of a number of symmetry principles and predictive properties. But subsequently the conclusions are reached by working out specific consequences for probability functions within this set, using the axioms of probability only. In particular, Carnapian inductive logic looks at the probability assignments conditional on various samples s_t, deriving by means of the probability axioms that they all satisfy the instantial relevance of Equation (4), for example. Importantly, in this template the symmetries in the language, like Equation (1) and Equation (2), appear as premises in the inductive logical inference. They restrict the set of probability assignments that is considered in the inference.

Carnapian logic and statistical inference are clearly similar, because they both concern sets of probability functions over sample space. However, while statistics frames these probability functions in terms of statistical hypotheses, these hypotheses do not appear in Carnapian logic. Instead, the emphasis is on characterising probability functions in terms of symmetries and predictive properties. The background of this is logical empiricism: the symmetries directly relate to the empirical predicates in the language of inductive logic, and the predictive properties relate to properties of the probability functions that show up for finite data. By contrast, statistical hypotheses are rather elusive: they cannot be formulated in terms of finite combinations of empirical predicates, and they only show up, if ever, in the limit of the data size going to infinity.

The overview of Carnapian logics given here is admittedly very brief. For example, I have not dealt with a notable exception to the *horror hypothesi* of inductive logicians, namely Hintikka systems. For more on the rich research programme of Carnapian inductive logic, I refer to chapter 9, and for Hintikka systems in particular, to chapter 10. For present purposes the thing to remember is that Carnapian logic can be viewed as a non-ampliative inductive logic, and that it does not make use of statistical hypotheses.

5 Bayesian statistics

The foregoing introduced Carnapian inductive logic. Now we can start answering the central question of this chapter. Can inductive logic accommodate statistical procedures?

The first statistical procedure under scrutiny is Bayesian statistics. The defining characteristic of this kind of statistics is that probability assignments do not just range over data, but that they can also take statistical hypotheses as arguments. As will be seen in the following, Bayesian inference is therefore naturally represented in terms of a non-ampliative inductive logic. Moreover, it relates very naturally to Carnapian inductive logic.

Let $\mathcal{H}$ be the space of statistical hypotheses h_θ, and let $\mathcal{Q}$ be the sample space as before. The functions P are probability assignments over the entire space $\mathcal{H} \times \mathcal{Q}$. Since the hypotheses h_θ are members of the combined algebra, the functions $P(s^t|h_\theta)$ range over the entire algebra $\mathcal{Q}$. We can define Bayesian statistics as follows.

Definition 5.1 (Bayesian Statistical Inference) *Assume the prior probability $P(h_\theta)$ assigned to hypotheses $h_\theta \in \mathcal{H}$, with $\theta \in \Theta$, the space of parameter values. Further assume $P(s_t|h_\theta)$, the probability assigned to the data s_t conditional on the hypotheses, called the likelihoods. Bayes' theorem determines that*

$$P(h_\theta|s_t) = P(h_\theta)\frac{P(s_t|h_\theta)}{P(s_t)}. \tag{5}$$

Bayesian statistics outputs the posterior probability assignment, $P(h_\theta|s_t)$.

See Barnett (1999) and Press (2003) for more detail. The further results form a Bayesian inference, such as estimations and measures for the accuracy of the estimations, can all be derived from the posterior distribution over the statistical hypotheses.

In this definition the probability of the data $P(s_t)$ is not presupposed, because it can be computed from the prior and the likelihoods by the law of total probability,

$$P(s_t) = \int_{\Theta} P(h_\theta) P(s_t|h_\theta) d\theta.$$

The result of a Bayesian statistical inference is not always a complete posterior probability. Often the interest is only in comparing the ratio of the posteriors of two hypotheses. By Bayes' theorem we have

$$\frac{P(h_\theta|s_t)}{P(h_{\theta'}|s_t)} = \frac{P(h_\theta)P(s_t|h_\theta)}{P(h_{\theta'})P(s_t|h_{\theta'})},$$

and if we assume equal priors $P(h_\theta) = P(h_{\theta'})$, we can use the ratio of the likelihoods of the hypotheses, the so-called Bayes factor, to compare the hypotheses.

Let me give an example of a Bayesian procedure. Say that we are interested in the colour composition of pears from Emma's farm, and that her pears are red, q_i^0, or green, q_i^1. Any ratio between these two kinds of pears is possible, so we have a set of hypotheses h_θ, called multinomial hypotheses, for which

$$P_{h_\theta}(q_t^1|s_{t-1}) = \theta, \qquad P_{h_\theta}(q_t^0|s_{t-1}) = 1 - \theta \tag{6}$$

with $\theta \in [0, 1]$. The hypothesis h_θ fixes the portion of green pears at θ, and therefore, independently of what pears we saw before, the probability that a randomly drawn pear from Emma's farm is green is θ. The type of distribution over $\mathcal{Q}$ that is induced by these hypotheses is sometimes called a Bernoulli distribution, or a multinomial distribution.

Let us define a Bayesian statistical inference over these hypotheses. Instead of straightforwardly choosing among them on the basis of the data, as classical statistics advises, we assign a probability density function over the range of hypotheses,

$$P(h_\theta) \propto \theta^{\lambda/2-1}(1-\theta)^{\lambda/2-1} \tag{7}$$

with $\theta \in \Theta = [0, 1]$. For $\lambda = 2$, this function is uniform over the domain. Now say that we observe a sequence of pears $s_t = s^{k_1 \ldots k_t}$. Defining t_1 as the number of green pears, or 1's, in the sequence s_t, and t_0 for the number of 0's, so $t_0 + t_1 = t$. The probability of these data given the hypothesis h_θ is

$$P(s_t|h_\theta) = \prod_{i=1}^{t} P_{h_\theta}(q_i^{k_i}|s_{i-1}) = \theta^{t_1}(1-\theta)^{t_0}. \tag{8}$$

Note that the probability of the data only depends on the number of 0's and the number of 1's in the sequence. Applying Bayes' theorem then yields:

$$P(h_\theta|s_t) \propto \theta^{\lambda/2-1+t_1}(1-\theta)^{\lambda/2-1+t_0}. \tag{9}$$

This is the posterior distribution over the hypotheses. It is derived by the axioms of probability theory alone, specifically by Bayes' theorem.

Most of the controversy over the Bayesian method concerns the determination and interpretation of the probability assignment over hypotheses. As will become apparent in the following, classical statistics objects to the whole idea of assigning probabilities to hypotheses. The data have a well-defined probability, because they consist of repeatable events, and so we can interpret the probabilities as frequencies, or as some other kind of objective probability. But the probability assigned to a hypothesis cannot be understood in this way, and instead expresses an epistemic state of uncertainty. One of the distinctive features of classical statistics is that it rejects such an epistemic interpretation of the probability assignment, and that it restricts itself a straightforward interpretation of probability as relative frequency.

Even if we buy into this interpretation of probability as epistemic uncertainty, how do we determine a prior probability? At the outset we do not have any idea of which hypothesis is right, or even which hypothesis is a good candidate. So how are we supposed to assign a prior probability to the hypotheses? The literature proposes several objective criteria for filling in the priors, for instance by maximum entropy or by other versions of the principle of indifference, but something of the subjectivity of the starting point remains. The strength of the classical statistical procedures is that they do not need any such subjective prior probability.

6 Inductive logic with hypotheses

Bayesian statistics is closely related to the inductive logic of Carnap. In this section I will elaborate on this relation, and indicate how Bayesian statistical inference and inductive logic may have a fruitful common future.

To see how Bayesian statistics and Carnapian inductive logic hang together, note first that the result of a Bayesian statistical inference, namely a posterior, is naturally translated into the result of a Carnapian inductive logic, namely a prediction,

$$P(q^1_{t+1}|s_t) = \int_0^1 P(q^1_{t+1}|h_\theta \cap s_t)P(h_\theta|s_t)d\theta, \tag{10}$$

by the law of total probability. Furthermore, consider the posterior probability over multinomial hypotheses. Recall that the parameter θ is the probability for the next pear to be green, see Equation (6). Therefore we have

$$\begin{aligned}
\text{Exp}[\theta] &= \int_\Theta \theta P(h_\theta|s_t)d\theta \\
&= \int_0^1 P(q^1_{t+1}|h_\theta \cap s_t)P(h_\theta|s_t)d\theta \\
&= P(q^1_{t+1}|s_t),
\end{aligned}$$

This shows that in the case of multinomial statistical hypotheses, the expectation value for the parameter is the same as a predictive probability.

But the correspondence becomes even more striking. We can work out the integral of Equation line (10), using Equation (9) as the posterior, to obtain

$$P(q^1_{t+1}|s_t) = \frac{t_1 + \lambda/2}{t + \lambda}. \tag{11}$$

This means that there is a specific correspondence between certain kinds of predictive probabilities, as described by the Carnapian λ-rules, and certain kinds of Bayesian statistical inferences, namely with multinomial hypotheses and priors of a particular shape.

The equivalence is in fact more general than this. Instead of the well-behaved priors just considered, we might consider as prior any functional form over the hypotheses h_θ, and then wonder what the resulting predictive probability is. As de Finetti (1937) showed in his representation theorem, the resulting predictive probability will always comply to a predictive property known as exchangeability, which was given in Equation (2). Conversely, and more surprisingly, any predictive probability complying to the property of exchangeability can be written down in terms of a Bayesian statistical inference with multinomial hypotheses and some prior over these hypotheses. In sum, de Finetti showed that there is a one-to-one correspondence between the predictive property of exchangeability on the one hand, and Bayesian statistical inferences using multinomial hypotheses on the other.

It is useful to make this result by de Finetti explicit in terms of the non-ampliative inductive logic discussed in the foregoing. Recall that a Bayesian statistical inference takes a prior and likelihoods as premises, leading to a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ as the only assignment satisfying the premises. We infer probabilistic consequences, such as the posterior and the predictions, from this probability assignment. Similarly, a Carnapian inductive logic is characterised by a single probability assignment, defined over the space $\mathcal{Q}$, from which the predictions can be derived. The representation theorem by de Finetti effectively shows an equivalence between these two probability assignments: when it comes to predictions, we can reduce the probability assignment over $\mathcal{H} \times \mathcal{Q}$ to an assignment over $\mathcal{Q}$ only.

For de Finetti, this equivalence was very welcome. He had a strictly subjectivist interpretation of probability, believing that probability expresses uncertain belief only. Moreover, he was eager to rid science of its metaphysical excess baggage to which, in his view, the notion of objective chance belonged. Therefore, in line with the logical empiricists working in inductive logic, de Finetti applied his representation theorem to argue against the use of multinomial hypotheses, and thereby against the use of statistical hypotheses more generally. Why refer to these obscure chances if we can achieve the very same statistical ends by employing the unproblematic notion of exchangeability? The latter is a predictive property, and it can therefore be interpreted as an empirical and as a subjective notion.

The fact is that statistics, as it is used in the sciences, is persistent in its use of statistical hypotheses. Therefore I want to invite the reader to consider the inverse application of de Finetti's theorem. Why does science use these obscure objective chances? As argued in Romeijn (2004, 2005), the reason is that statistical hypotheses provide invaluable help by, indirectly, pinning down the probability assignments over $\mathcal{Q}$ that have the required predictive properties.

Rather than reducing the Bayesian inferences over statistical hypotheses to inductive predictions over observations, we can use the representation theorem to capture relations between observations in an insightful way, namely by citing the statistical hypotheses that may be true of the data. In terms of Carnapian inductive logic, we are thereby extending the language of inductive logic with theoretical terms. As illustrated in Romeijn (2006), enriching inductive logic in this way improves the control that we have over predictive properties. Hence it seems a rather natural extension of traditional Carnapian inductive logic.

Bayesian statistics, as it has been presented here, is a ready made specification of this extended inductive logic, which may be called *Bayesian inductive logic*. The premises of the inference are restrictions to the set of probability assignments over $\mathcal{H} \times \mathcal{Q}$, and the conclusions are simply the probabilistic consequences of these restrictions, derived by means of the axioms of probability, often Bayes' theorem. The inferential step, as in Carnapian logic, is thus non-ampliative. When it comes to the predictive consequences, the extension of the probability space with $\mathcal{H}$ may be considered unnecessary because, as indicated, we can always project the probability P over the extended space back onto $\mathcal{Q}$. However, the probability function resulting from that projection may be very hard to define in terms of its predictive properties alone.

Capturing Bayesian statistics in this inductive logic is immediate. The premises are the prior over the hypotheses, $P(h_\theta)$ for $\theta \in \Theta$, and the likelihood functions, $P(s_t|h_\theta)$ over the algebras $\mathcal{Q}$, which are determined for each hypothesis h_θ separately. These premises are such that only a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ remains. In other words, the premises have a unique probability model. The conclusions all follow from the posterior probability over the hypotheses. All these conclusions can be derived from the assignment by applying theorems of probability theory, primarily Bayes' theorem.

Finally, let me relate this view on inductive logic to a comparable view, expressed by Hintikka in Auxier and Hahn (2006). In response to Kuipers' overview of inductive logic, Hintikka writes that "Inductive inference, including rules of probabilistic induction, depends on tacit assumptions concerning the nature of the world. Once these assumptions are spelled out, inductive inference becomes in principle a species of deductive inference." The symmetry principles and predictive properties used in Carnapian inductive logic amount to such tacit assumptions. The use of particular statistical hypotheses in a Bayesian inference come down to the same, but here they are not tacit anymore. They can thus help to understand and control these assumptions.

7 Neyman-Pearson testing

In the foregoing, I have presented Carnapian inductive logic and Bayesian statistical inference. I have shown that these two are strongly related, and that they both fit the template of non-ampliative inductive logic introduced in section 3. This led to the introduction of Bayesian inductive logic in the preceding section. In the following, I will consider two classical statistical procedures, Neyman-Pearson hypothesis testing and Fisher's maximum likelihood estimation, and see whether they can be captured in this inductive logic.

Neyman-Pearson hypothesis testing concerns the choice between two statistical hypotheses, that is, two fully specified probability functions over sample space. Let $\mathcal{H} = \{h_0, h_1\}$ be the set of hypotheses, and let $\mathcal{Q}$ be the sample space, that is, the observation algebra introduced earlier on. Each of the hypotheses is associated with a complete probability function P_{h_j} over the sample space. But note that, unlike in Bayesian statistics, the hypotheses h_j are not part of the probability space, so no probability is assigned to the hypotheses themselves, and we cannot write $P(\cdot|h_j)$ anymore.

In Neyman-Pearson statistics we compare the hypotheses h_0 and h_1 by means of a so-called test function. See Barnett (1999) and Neyman and Pearson (1967) for the details.

Definition 7.1 (Neyman-Pearson Hypothesis Test) *Let F be a function over the sample space $\mathcal{Q}$,*

$$F(s_t) = \begin{cases} 1 & \textit{if } \frac{P_{h_1}(s_t)}{P_{h_0}(s_t)} > r, \\ 0 & \textit{otherwise,} \end{cases} \tag{12}$$

where P_{h_j} is the probability over the sample space determined by the statistical hypothesis h_j. If $F = 1$ we decide to reject the null hypothesis h_0, else we reject the alternative h_1.

Note that, in this simplified setting, the test function is defined for each set of sequences s_t separately. For each sample plan, and associated sample size t, we must define a separate test function.

The decision to accept or reject a hypothesis reject is associated with the so-called significance and power of the test:

$$\text{Significance}_F = \alpha = \int_{\mathcal{Q}} F(s_t) P_{h_0}(s_t) ds_t,$$

$$\text{Power}_F = 1 - \beta = \int_{\mathcal{Q}} F(s_t)) P_{h_1}(s_t) ds_t.$$

The significance is the probability, according to the hypothesis h_0, of obtaining data that leads us to reject the hypothesis h_0, or in short, the type-I error of falsely rejecting the null hypothesis, denoted α. Similarly, the power is the probability, according to h_1, of obtaining data that leads us to reject the hypothesis h_0, or in short, the probability of correctly rejecting the null hypothesis, so that $\beta = 1 - \text{Power}$ is the type-II error of falsely accepting the null hypothesis. An optimal test is one that minimizes the significance level, and maximizes the power. Neyman and Pearson prove that the decision has optimal significance and power for, and only for, likelihood-ratio test functions F. That is, an optimal test depends only on a threshold for the ratio $\frac{P_{h_1}(s_t)}{P_{h_0}(s_t)}$.

Let me illustrate the idea of Neyman-Pearson tests. Say that we have a pear whose colour is described by q^k, and we want to know from what farm it originates, from farmer Maria (h_0) or Lisa (h_1). We know that the colour composition of the pears from the two farms are as follows:

Hypothesis \ Data	q^0	q^1	q^2
h_0	0.00	0.05	0.95
h_1	0.40	0.30	0.30

If we want to decide between the two hypotheses, we need to fix a test function. Say that we choose

$$F(q^k) = \begin{cases} 0 & \text{if } k = 2, \\ 1 & \text{else.} \end{cases}$$

In the definition above, which uses a threshold for the likelihood ratio, this comes down to choosing a value for r somewhere between $6/19$ and 14, for example $r = 1$. The significance level is $P_{h_0}(q^0 \cup q^1) = 0.05$, and the power is $P_{h_1}(q^0 \cup q^1) = 0.70$. Now say that the pear we have is green, so $F = 1$ and we reject the null hypothesis, concluding that Maria did not grow the pear with the aforementioned power and significance.

Note that from the perspective of ampliative inductive logic, it is not too far-fetched to read an inferential step into the Neyman-Pearson procedure. The test function F brings us from a sample s_t and two probability functions, P_{h_j} for $j = 0, 1$, to a single probability function P_{h_j} over the sample space $\mathcal{Q}$. So we might say that the test function is the procedural analogue of an inductive inferential step, as discussed in Section 3. This step is ampliative because both probability functions P_{h_j} are consistent with the data. Ruling out one of them cannot be done deductively.

There are attempts to make these ampliative inferences more precise, by means of a form of default reasoning, or a reasoning that employs a preferential ordering over probability models. Specifically, so-called evidential probability, proposed by Kyburg (1974) and more recently discussed by Wheeler (2006), is concerned with inferences that combine statistical hypotheses, which are each accepted with certain significance levels. However, in this chapter I will not investigate these logics. They are not concerned with inferences from the data to predictions or to hypotheses, but rather with inferences from hypotheses to other hypotheses, and from hypotheses to predictions.

Neyman-Pearson hypothesis testing is sometimes criticised because its results depend on the shape of the entire sample space, and not just on the observed sample. That is, the decision to accept or reject the null hypothesis against some alternative hypothesis depends not just on what has actually been observed in experiment, but also on what could have been observed. A well-known illustration of this concerns so-called optional stopping. But here I want to illustrate the same point with an example due to Jeffreys (1931), which is discussed at length in Hacking (1965). I hope it helps to understand Neyman-Pearson statistics better.

Instead of the hypotheses h_0 and h_1 above, say that we compare the hypotheses h'_0 and h_1.

Hypothesis \ Data	q^0	q^1	q^2
h'_0	0.05	0.05	0.90
h_1	0.40	0.30	0.30

We determine the test function $F(q^k) = 1$ iff $k = 0$, by requiring the same significance level, $P_{h'_0}(q^0) = 0.05$, resulting in the power $P_{h_1}(q^0) = 0.40$. Now imagine that we observe q^1 again, so that we accept h'_0. But this is a bit odd, because the hypotheses h_0 and $h^\star_0$ have the same probability for q^1! So how can they react differently to this observation? It seems that, in contrast to h_0, the hypothesis h'_0 escapes rejection because it allocates some probability to s^0, an event that does not occur, thus shifting the area in sample space on which it is rejected. The example gave rise to the famed complaint of Jeffreys that "the null hypothesis can be rejected because it fails to predict an event that never occurred".

This illustrates how the results of a Neyman-Pearson procedure depends on the shape of the sample space, and not just on the actual observation. From the perspective of an inductive logician, it may therefore seem "a remarkable procedure", to cite Jeffreys again. But it must be emphasised that Neyman-Pearson statistics was never intended as an inference in disguise. It is a procedure that allows us to decide between two hypotheses on the basis of data, generating error rates associated with that decision. Neyman and Pearson themselves were very explicit that the procedure must not be interpreted inferentially. Rather than inquiring into the truth and falsity of a hypothesis, they were interested in the probability of mistakenly deciding to reject or accept a hypothesis. The significance and power concern the probability over data given a hypothesis, not the probability of hypotheses given the data.

8 Neyman-Pearson test as an inference

In this section, I investigate whether we can turn the Neyman-Pearson procedure of Section 7 into an inference within Bayesian inductive logic. This might come across as a pointless exercise in statistical yoga, trying to make Neyman and Pearson relax in a position that is far from natural. However, the exercise will nicely illustrate how statistics may be related to inductive logic, and thereby stimulate research on the intersection of inductive logic and statistics in the sciences.

An additional reason for investigating Neyman-Pearson hypothesis testing in this framework is that in many practical applications, scientists are tempted to read the probability statements about the hypotheses inversely after all: the significance is often taken as the probability that the null hypothesis is true. Although emphatically wrong, this inferential reading has a strong intuitive appeal to users. The following will make explicit that in this reading, the Neyman-Pearson procedure is effectively taken as a non-ampliative entailment.

To this aim, we construct the space $\mathcal{H} \times \mathcal{Q}$, and define the probability functions P_{h_j} over the sample spaces $\langle h_j, \mathcal{Q} \rangle$. For the prior probability assignment over the two hypotheses, we take $P(h_0) \in (l, u)$, meaning that $l < P(h_0) < u$. We write $\underline{P}(h_j) = \min P(h_j)$ and $\overline{P}(h_j) = \max P(h_j)$. Finally, we adopt the restriction that $P(h_0) + P(h_1) = 1$. This defines a set of probability functions over the entire space, serving as a starting point of the inference.

Next we include the data in the probability assignments. Crucially, I coarse-grain the observations to the simple observation f^j, with

$$f^j = \{s_t : F(s_t) = j\},$$

so that the observation simply encodes the value of the test function. We then have the type-I and type-II errors as the likelihoods of the observations,

$$\begin{aligned} P(f^1|h_0) &= \alpha, \\ P(f^0|h_1) &= \beta. \end{aligned}$$

Finally we use Bayes' theorem to derive a set of posterior probability distributions over the hypotheses, according to

$$\frac{P(h_1|f^j)}{P(h_0|f^j)} = \frac{P(f^j|h_1)P(h_1)}{P(f^j|h_0)P(h_0)}.$$

Note that the quality of the test, in terms of size and power, will be reflected in the posteriors. If, for example, we find an observation s_t that allows us to reject the null hypothesis, so f^1, then for the posterior interval we will generally have $\underline{P}(h_0|f^1) < \underline{P}(h_0)$ and $\overline{P}(h_0|f^1) < \overline{P}(h_0)$.

With this representation, we have not yet decided on a fully specified prior probability over the statistical hypotheses. This echoes the fact that classical statistics does not make use of a prior probability. However, it is only by restricting the prior probability over hypotheses in some way or other that we can make the Bayesian rendering of the results of Neyman and Pearson work. In particular, if we choose $(l, u) = (0, 1)$ for the prior, then we find the interval $(0, 1)$ for the posterior as well. However, if we choose

$$l \geq \frac{\beta}{\beta + 1 - \alpha}, \qquad u \leq \frac{1 - \beta}{1 - \beta + \alpha},$$

we find for all $P(h^0) \in (l, u)$ that $\overline{P}(h_0|f^1) < 1/2 < \underline{P}(h_1|f^1)$. Similarly, we find $\underline{P}(h_0|f^0) > 1/2 > \overline{P}(h_1|f^0)$. So with this interval prior, an observation s_t for which $F(s_t) = 1$ tilts the balance towards h_1 for all the probability functions P in the interval, and vice versa.

Let me illustrate the Bayesian inference by means of the above example on pears. We set up the sample space and hypotheses as before, and we then coarse-grain the observations to f^j, corresponding to the value of the test function, $f^1 = q^0 \cup q^1$ and $f^0 = q^2$. We obtain

$$\begin{aligned} P(f^1|h_0) &= P(q^0 \cup q^1|h_0) = \alpha = 0.05 \\ P(f^0|h_1) &= P(q^0 \cup q^1|h_1) = \beta = 0.30 \end{aligned}$$

Choosing $P(h_0) \in (0.24, 0.93)$, this results in $P(h_0|f^0) = (0.50, 0.98)$, and $P(h_0|f^1) = (0.02, 0.50)$.

Depending on the choice of prior, we might claim that the resulting Bayesian inference replicates the Neyman-Pearson procedure: if the probability over hypotheses expresses our preference over them, then indeed f^0 makes us prefer h_0 and f^1 makes us prefer h_1. Moreover, the inference fits the entailment relation mentioned earlier: we have a set of probabilistic models on the side of the premises, namely the set of priors over $\mathcal{H}$, coupled to the full probability assignments over $\langle h_j, \mathcal{Q} \rangle$ for $j = 0, 1$. And we have a set of models on the conclusion side, namely the set of posteriors over $\mathcal{H}$. Because the latter is computed from the former by the axioms of probability, the two sets are identical. Therefore the conclusion is classically entailed by the premises.

On the other hand, the results of a Bayesian inference will always be a probability function. By contrast, Neyman-Pearson statistics ends in a decision to accept or reject, which is binary instead of some sort of weak preference. That bivalence cannot be replicated in a Bayesian rendering. We may supplement the probabilistic results of a Bayesian inference with rules for translating the probability assignments into decisions. As suggested, we might say that we choose h_0 if we have $\underline{P}(h_0|s_t) > 1/2$, and similarly for h_1. But as attested by the vast literature on the lottery paradox, this quickly leads to problematic consequences. On this count, the portrayal of Neyman-Pearson statistics as an inference simply does not work.

More in general, the representation in probabilistic logic will probably not appeal to advocates of classical statistics. Quite apart from the issue of binary acceptance, the whole idea of assuming a prior probability, however unspecific, may be objected to on the principled ground that probability functions express long-term frequencies, and that hypotheses cannot have such frequencies.

There is one attractive feature, at least to my mind, of the above rendering, that may be of interest in its own right. With the representation in place, we can ask again how to understand the example by Jeffreys. For Hacking it illustrates that Neyman and Pearson do not respect the likelihood principle. However, in the above representation, we do respect the likelihood principle. Instead, we could say that the approach of Neyman and Pearson takes the observations to present rather coarse-grained information. Specifically, in the Bayesian representation we condition on f^j and not on q^k. The whole example hinges on how the samples are grouped into regions of acceptance and rejection. Within this viewpoint, therefore, the sample space dependence of Neyman-Pearson procedures may be taken as an illustration of the idea that the content of observation depends on how the observation is framed.

9 Fisher's parameter estimation

Let me turn to another important classical statistical procedure, so-called parameter estimation. I focus in particular on an estimation procedure first devised by Fisher, namely maximum likelihood estimation. The two sections following this one will be devoted to the question if and how we can capture this classical statistical procedure in Bayesian inductive logic.

The maximum likelihood estimator determines the best among a much larger, possibly infinite, set of hypotheses. It depends on the probability that the hypotheses assign to points in the sample space. See Fisher (1956) and Barnett (1999) for the details.

Definition 9.1 (Maximum Likelihood Estimation) *Let $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ be a set of hypotheses, labeled by the parameter θ, and let $\mathcal{Q}$ be the* *observation algebra. Then the maximum likelihood estimator of* θ*,*

$$\hat{\theta}(s_t) = \{\theta : \forall h_{\theta'} \big(P_{h_{\theta'}}(s_t) \leq P_{h_\theta}(s_t) \big) \}, \tag{13}$$

is a function over the elements s_t *in the sample space.*

So the estimator is a set, typically a singleton, of those values of θ for which the likelihood of h_θ on the data s_t is maximal. The associated best hypothesis we denote with $h_{\hat{\theta}(s_t)}$, or $h_{\hat{\theta}}$ for short. The estimator is a function over the sample space, associating each s_t with a hypothesis, or a set of them.

Often the estimation is coupled to a so-called confidence interval. Restricting the parameter space to $\Theta = [0, 1]$ for convenience, and assuming that the true value is θ, we can define a region in sample space within which the estimator function is not too far off the mark. Specifically, we might set the region in such a way that it covers $1 - \epsilon$ of the probability P_{h_θ} over sample space:

$$\int_{\theta-\Delta}^{\theta+\Delta} P_{h_\theta}(\hat{\theta}) d\hat{\theta} = 1 - \epsilon \tag{14}$$

We can provide an unproblematic frequentist interpretation of the interval $\hat{\theta} \in [\theta - \Delta, \theta + \Delta]$. In a series of estimations, the fraction of times in which the estimator $\hat{\theta}$ is further off the mark than Δ will tend to ϵ. The smaller the region, the more reliable the estimate. Note, however, that this interval is defined in terms of the unknown true value θ. In Section 11, I will introduce an alternative notion of confidence interval that avoids this drawback.

For now, let me illustrate parameter estimation in a simple example on pears, concerning the statistical hypotheses defined in Equation (6). Now the general idea is that we choose the value of θ for which the probability that the hypothesis gives to the data is maximal. Recall that the likelihoods of the multinomial hypotheses h_θ are

$$\theta^{t_1}(1 - \theta)^{t_0}.$$

This function is maximal at $\theta = {}^{t_1}/_{t}$, so the maximum likelihood estimator is

$$\hat{\theta}(s_t) = \frac{t_1}{t}. \tag{15}$$

Note finally that for a true value θ, the probability of finding the estimate in the confidence interval of Equation (14),

$$\frac{t_1}{t} \in [\theta - \Delta, \theta + \Delta]$$

increases for larger data sequences. Fixing the probability at $1-\epsilon$, the size of the interval will therefore decrease.

This completes the introduction into parameter estimation. Note that the statistical procedure can be taken as the procedural analogue of an ampliative logical inference, running from the data to a probability assignment over the sample space. We have $\mathcal{H}$ as the set of probability assignments or hypotheses from which the inference starts, and by means of the data we then choose a single $h_{\hat{\theta}}$ from these as our conclusion. However, in the following I aim to investigate whether there is a non-ampliative logical representation of this inductive inference.

10 Estimations in inductive logic

There are at least two ways in which parameter estimation can be turned into a non-ampliative logic. One of these, fiducial inference, generates a probability assignment over statistical hypotheses without presupposing a prior probability at the outset. We deal with this inference in the next section. In this section, we investigate the relation between parameter estimation and the non-ampliative inductive logics devised in the foregoing.

To spot the similarity between parameter estimation and Carnapian inductive logic, note that the procedure of parameter estimation can be used to determine the probability of the next piece of data. In the example on pears, once we have observed s^{000101}, say, we choose $h_{1/3}$ as our best estimate, and we may on the basis of that predict that the next pear has a probability of $1/3$ to be green. The function $\hat{\theta}$ is then used as a predictive system, much like any other Carnapian inductive logic:

$$P(q^k_{t+1}|s_t) = P_{h_{\hat{\theta}(s_t)}}(q^k_{t+1}).$$

The estimation function $\hat{\theta}$ by Fisher is thereby captured in a single probability function P. So we can present the latter as a probability assignment over sample space, from which estimations can be derived by a non-ampliative inference.

Let me make this concrete by means of the example on red and green pears. In the Carnapian prediction rule of Equation (3), choosing $\lambda = 0$ will yield the observed relative frequency as predictions. And according to Equation (15) these relative frequencies are also the maximum likelihood estimators. Thus, for each set of possible observations, $\{s^{k_1 \dots k_t} : k_i = 0, 1\}$, the Carnapian rule with $\lambda = 0$ predicts according to the Fisherian estimate.

Note that the probability function P that describes the estimations is a rather unusual one. After three red pears for example, s^{000}, the probability for the next pear to be green will be 0, so that $P(s^{0001}) = 0$. Then, by the standard axiomatisation and definitions of probability, the probability of any observation q_5^0 conditional on s^{0001} is not defined. But if the probability function P is supposed to follow the Fisherian estimations, then we must have $P(q_5^0 | s^{0001}) = 3/4$. To accommodate the probability function imposed by Fisher's estimations, we must therefore change the axiomatisation of probability. In particular, we may adopt an axiomatisation in which conditional probability is primitive, as described in Rényi (1970) and in chapter 15 of this volume.

Apart from the nonstandard axiomatisation, the alignment of Fisher estimation and Carnapian inductive logic is not exactly easy. For more complicated sets of hypotheses, and the more complicated estimators associated with it, the corresponding probability assignment P may be even less natural. Moreover, the principles and predictive properties that motivate the choice of that probability function will be very hard to come by. In the following I will therefore not discuss the further intricacies of capturing Fisher's estimation functions by Carnapian prediction rules.

Instead, I want to devote some attention to capturing parameter estimation in Bayesian statistical inference, and thereby in inductive logic with hypotheses. Note that in both parameter estimation and Bayesian statistics, we consider a set of statistical hypotheses and we are looking to find the best fitting one. Moreover, in both of these our choice among these is informed by the probability of the data according to the hypotheses. Bayesian inductive logic, the non-ampliative inductive logic that emulates Bayesian statistics, is therefore more suitable for capturing parameter estimation than Carnapian inductive logic.

To capture something like parameter estimation, note that the posterior over hypotheses can be used to generate the kind of choices between hypotheses that classical statistics provides. As for parameter estimation, note that we can use the posterior to derive an expectation for the parameter θ, as follows:

$$\mathrm{E}[\theta] = \int_{\Theta} \theta P(h_\theta | s_t) d\theta.$$

Clearly, $\mathrm{E}[\theta]$ is a function that brings us from the hypotheses h_θ and the data s_t to a preferred value for the parameter. The function depends on the prior probability over the hypotheses, but it is nevertheless analogous to the maximum likelihood estimator.

In analogy to the confidence interval, we can also define a so-called credal interval from the posterior probability distribution:

$$\mathrm{Cred}_{1-\epsilon}(s_t) = \left\{ \theta : \; |\theta - \mathrm{E}[\theta]| < \Delta \text{ and } \int_{\mathrm{E}[\theta]-\Delta}^{\mathrm{E}[\theta]+\Delta} P(h_\theta | s_t) d\theta = 1 - \epsilon \right\}.$$

Therefore,

$$P(\{\theta : \; \theta \in \mathrm{Cred}_{1-\epsilon}(s_t)\} | s_t) = 1 - \epsilon. \qquad (16)$$

This set of values for θ is such that the posterior probability of the corresponding h_θ jointly add up to $1 - \epsilon$ of the total posterior probability. We might argue that this expression is an improvement over the classical confidence interval of Equation (14). The latter only expresses how far an estimate is off the mark on average, while it does not warrant an inference about how far away the specific estimate that we have obtained, lies with respect to the true value of the parameter. By contrast, a credal interval does allow for such an inferential reading.

Of course there are also large differences between the results of parameter estimation and the results of a Bayesian analysis. One difference is that in parameter estimation, and in classical statistics more generally, the choice for some hypothesis is an all-or-nothing affair: we accept or reject, we choose a single best estimate, and so on. In the Bayesian procedure, by contrast, the choice is expressed in a posterior probability assignment over the set of hypotheses. As indicated in the discussion of Neyman-Pearson hypothesis testing, this difference remains problematic.

In addition, there is a well-known, but no less grave drawback to the way in which the Bayesian conclusions are reached: we have to assume a prior probability assignment over the statistical hypotheses. Any expectation and credal interval depends on the exact prior that is chosen. This dependence can only be avoided by assuming that we have sufficient data to swamp the impact of the prior or, equivalently, by assuming that the prior is sufficiently smooth in comparison to the likelihoods for the data.

11 Fiducial probability

This latter problem, of how to choose the prior, motivated Fisher (1930, 1935, 1956) to devise an alternative way of making parameter estimation inferential, by means of the so-called fiducial argument. This argument yields a probability assignment over hypotheses without assuming a prior probability over statistical hypotheses at the outset. The fiducial argument is controversial, however, and its applicability is limited to particular statistical problems. See Hacking (1965) and Seidenfeld (1979) for detailed critical discussions, and Barnett (1999) for a good overview. In the following, I will only provide a brief sketch of the argument.

A good way of introducing fiducial probability is by the notion of confidence intervals, introduced in Section 9. In some cases, we can also derive a region of parameter values within which the true value θ can be expected to lie. The general idea is to define a set of parameter values R within which the data are not too unlikely, $R(s_t) = \{\theta : P_{h_\theta}(s_t) > 1\%\}$. Specifically, in terms of the integral in Equation (14), we can swap the roles of θ and $\hat{\theta}$ and define:

$$\mathrm{Fid}_{1-\epsilon}(s_t) = \left\{ \theta : |\theta - \hat{\theta}(s_t)| < \Delta \text{ , and } \int_{\hat{\theta}-\Delta}^{\hat{\theta}+\Delta} P_{h_\theta}(\hat{\theta}) d\theta = 1 - \epsilon \right\}. \tag{17}$$

Every element of the sample space s_t is thus assigned a so-called fiducial interval $\mathrm{Fid}_{1-\epsilon}$ of parameter values.

Note, however, that the integral of Equation (17) only properly concerns a probability if we have

$$P_{h_\theta}(\hat{\theta} + \delta) = P_{h_{\theta-\delta}}(\hat{\theta})$$

for all values of δ. If that condition is met, then for any fixed value of $\hat{\theta}$, the function $P_{h_\theta}(\hat{\theta})$ is indeed a probability density over the variable θ. In that case, we can interpret this interval in much the same way as the credal interval of Equation (16), as a probability:

$$P(\{\theta : \theta \in \mathrm{Fid}_{1-\epsilon}(s_t)\}|s_t) = 1 - \epsilon. \tag{18}$$

But if the condition is not met, the interval cannot be taken as expressing a probability that the true value of the parameter lies within a certain interval around the estimate. Or at least, we cannot interpret it in this way without further consideration.

The determination of the intervals of Equation (17) is an example of the determination of fiducial probability. But note that it relies on a strong requirement. We must presuppose the equivalence of two distinct functions, both written $P_{h_\theta}(\hat{\theta})$, one taking θ and one taking $\hat{\theta}$ as argument. A much more general formulation of this requirement is provided by Dawid and Stone (1982). They argue that in order to run the fiducial argument, one has to assume that the statistical problem can be captured in a functional model that is smoothly invertible.

I want to conclude with an explanation of the notion of a smoothly invertible functional model. Assume that there is some function of the data $\hat{\theta}(s_t)$ relating the statistical hypothesis h_θ and an error term ω according to

$$\hat{\theta}(s_t) = f(h_\theta, \omega).$$

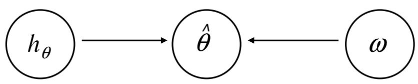
Now we assume a probability function $P(\omega)$ over the error terms, so that the functional relation and the probability over error terms together determine a probability

$$P(\hat{\theta}(s_t)|h_\theta) = P(\{\omega : \ f(\theta, \omega) = \hat{\theta}\}). \tag{19}$$

Suppose that the function f is invertible: we also have a function $f^{-1}(\hat{\theta}, \omega) = \theta$. And finally, we assume that the error terms and the hypotheses are probabilistically independent:

$$P(h_\theta, \omega) = P(h_\theta)P(\omega). \tag{20}$$

As described in Neapolitan (2003), for example, we can write down the overall probability assignment in terms of a graphical structure, as depicted below.



Say that we observe s_t, thus fixing the value for $\hat{\theta}(s_t)$, and that we condition on this observed data. Then, because of the network structure and the further fact that the relation $f(h_\theta, \omega)$ is deterministic, the variables ω and h_θ become perfectly correlated: each ω is associated with a unique $\theta = f^{-1}(\hat{\theta}, \omega)$. And because the observation of s_t does not itself influence the probability of ω either, meaning that $P(\omega|s_t) = P(\omega)$, we can write

$$P(h_\theta|\hat{\theta}(s_t)) = P(\{\omega : \ f^{-1}(\hat{\theta}, \omega) = \theta\}), \tag{21}$$

which is the inverse of Equation (19). This means that after observing s_t we can transfer the probability distribution over ω onto h_θ according to the function f^{-1}.

The fiducial probability over the hypotheses h_θ is, I think, a surprising result. No prior probability has been assumed, and nevertheless the construction is such that we can derive something that looks like a posterior. Moreover, the inductive inference in this construction is non-ampliative. The set of probability assignments over h_θ, s_t, and ω is such that $P(h_\theta)$ can be any convex combination of elements from a set of functions over θ, while $P(h_\theta|s_t)$ is a distinct element from that set. Given the controversy that surrounds the interpretation and determination of prior probabilities, it is a real pity that the fiducial argument can only be run under such strict conditions.

12 In conclusion

In the foregoing I have introduced a setting in which inductive logic and statistics may be unified. I have discussed how inductive logic can be developed to encompass and emulate a number of inductive procedures from mathematical statistics. In particular, the discussion of Bayesian statistical inference has led to the extension of the language of inductive logic with statistical hypotheses. The resulting inductive logic was applied to two classical procedures, to wit, Neyman-Pearson hypotheses testing and Fisher's maximum likelihood estimation. While these procedures are best understood as ampliative inductive inferences, I have shown that they can also be modelled, at least partly, in terms of this extended inductive logic.

I hope that portraying statistical procedures in the setting of non-ampliative inductive logic has been illuminating. More than that, I hope that the relation between Carnapian inductive logic and Bayesian and classical statistics stimulates research on the intersection of the two. For example, it will be nice to know, from the perspective of inductive logic, what kind of terms these elusive statistical hypotheses are. I believe that an inductive logic that includes statistical hypotheses in its language is very closely related to statistics. Research on their intersection can greatly enhance the relevance of inductive logic to the philosophy of science, and ultimately to science itself.

Acknowledgements

I want to thank the Dutch Organization of Scientific Research (NWO) for generous support (VENI-grant nr. ...).

References

Auxier, R. and Hahn, L., editors (2006). *The Philosophy of Jaako Hintikka*. Open Court, Chicago.

Bandyopadhyay, P. and Forster, M., editors (2009). *Handbook for the Philosophy of Science: Philosophy of Statistics*. Elsevier.

Barnett, V. (1999). *Comparative Statistical Inference*. John Wiley, New York.

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press.

Carnap, R. (1952). *The Continuum of Inductive Methods*. University of Chicago Press, Chicago.

Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference (with discussion). *Annals of Statistics*, 10(4):1054–1074.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68.

Festa, R. (2006). Analogy and exchangeability in predictive inferences. *Erkenntnis*, 45:89–112.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6:317–324.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

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

Haenni, R., Romeijn, J.-W., Wheeler, G., and Williamson, J. (2008). Probabilistic logics and probabilistic networks. manuscript.

Hailperin, T. (1996). *Sentential Probability Logic.* Lehigh University Press.

Howson, C. (2003). Probability and logic. *Journal of Applied Logic*, 1(3–4):151–165.

Jeffreys, H. (1931). *Scientific Inference.* Cambridge University Press, , Cambridge.

Kyburg, Jr., H. E. (1974). *The Logical Foundations of Statistical Inference.* D. Reidel, Dordrecht.

Maher, P. (2000). Probabilities for two properties. *Erkenntnis*, 52:63–81.

Neapolitan, R. E. (2003). *Learning Bayesian Networks.* Prentice Hall.

Neyman, J. and Pearson, E. (1967). *Joint Statistical Papers.* University of California Press, Berkeley.

Paris, J. and Waterhouse, P. (2008). Atom exchangeability and instantial relevance. *unpublished manuscript.*

Press, J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications.* John Wiley, New York.

Rényi, A. (1970). *Probability Theory.* North Holland, Amsterdam.

Romeijn, J. (2004). Hypotheses and inductive predictions. *Synthese*, 141(3):333–64.

Romeijn, J. (2005). *Bayesian Inductive Logic.* PhD dissertation, University of Groningen.

Romeijn, J. (2006). Analogical predictions for explicit similarity. *Erkenntnis*, 64:253–280.

Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher.* Reidel, Dordrecht.

Wheeler, G. (2006). Rational acceptance and conjunctive/disjunctive absorption. *Journal of Logic, Language and Information*, 15(1-2):49–63.