Abducted by Bayesians?

Jan-Willem Romeijn
Department of Psychology,
University of Amsterdam
j.w.romeijn@uva.nl

1. Introduction. This paper discusses the role of theoretical notions in Bayesian statistical inference. If we say that the observational content of statistical hypotheses is given by their likelihoods, any distinction between hypotheses that is not reflected in a difference in likelihoods is based on a theoretical notion. Typical examples of such theoretical notions are simplicity and explanatory force: two hypotheses may predict the observations according to identical probability models, yet the one hypothesis may be simpler, or associated with a better story on the underlying mechanism, than the other. The aim of this paper is to elucidate the use of such theoretical notions. It will be argued that even within Bayesian statistical inference, such notions can play an active role. The arguments will lead to two claims. First, the underdetermination resulting from theoretical notions is better seen as a methodological tool than as a problem. And second, the use of theoretical notions in Bayesian statistical inference may be viewed as a Bayesian explication of abduction.

2. Statistical hypotheses. This section presents the Bayesian inference of predictions and parameter estimations, as defined in Howson and Urbach (1989) and Good (1955). Because the statistical hypotheses of this section differ in their likelihoods, the notion that distinguishes the hypotheses is, according to the foregoing, not theoretical. On the other hand, likelihoods are typically associated with the rather theoretical notion of chance. Moreover, using the representation

theorem by De Finetti (1937) it may even be argued that the hypotheses concerning these chances can and must be eliminated from the scheme. Against this view, I argue that hypotheses on chances serve a specific purpose. They express the chance mechanism that is supposed to underlie the observations, and additional knowledge of these mechanisms can be expressed conveniently in a prior probability over them.

Bayesian inference employs a formal framework of observations, as laid down in an observation algebra, and a space of hypotheses. Consider an observation at time i with a possible result $q \in \{0, 1\}$, denoted Q_i^q, and denote sequences of observations of length t with E_t. In the example of this paper, the observations are results of coin tosses. Consider a partition $\mathcal{H}$ of hypotheses H_θ concerning the chance θ on tails, $q = 1$, so that

$$p(Q_{i+1}^1 | H_\theta \cap E_i) = \theta, \tag{1}$$

and further a prior probability over these hypotheses, $p(H_\theta)d\theta$. The partition, the likelihoods of the hypotheses in it, and the prior probability over the hypotheses together determine the Bayesian inference.

The observations are the other input component that is needed to arrive at predictions and parameter estimations. Bayesian conditioning can be used to update the prior probability to the given observations E_t. After updating we obtain a posterior probability,

$$p(H_\theta | E_t)d\theta = \frac{p(E_t|H_\theta)}{p(E_t)} p(H_\theta)d\theta. \tag{2}$$

Predictions follow directly from this posterior by the law of total probability:

$$\begin{aligned} p(Q_{t+1}^1 | E_t) &= \int_0^1 p(Q_{t+1}^1 | H_\theta \cap E_t)\, p(H_\theta | E_t)\, d\theta \\ &= \int_0^1 \theta\, p(H_\theta | E_t)\, d\theta. \end{aligned} \tag{3}$$

This latter expression is the expectation value for θ after observing E_t, which may also be read as a parameter estimation. Note further that this scheme for predictions covers a substantial part of Bayesian statistical inference, because many such inferences are made with models concerning constant chances.

In a sense, the hypotheses H_θ are already theoretical. They concern the objective chance of an observation, and such chances cannot be translated into finite observational terms. Moreover, the hypotheses can be eliminated from the inference completely, and following Hintikka (1970) any real empiricist should in fact strive for that elimination . De Finetti's representation theorem states that the above scheme of hypotheses covers exactly those prediction rules, or estimation functions, for which the order of the observations in E_t is inessential. Defining t_q as the number of Q_i^q in E_t, these rules may be characterised with

$$p(Q_{t+1}^q|E_t) = pr(t_q, t). \tag{4}$$

Every rule pr corresponds to a specific prior $p(H_\theta)d\theta$ over the hypotheses in the scheme, and vice versa. In particular, following Carnap (1952) and Festa (1993), if we assume the prior to be a symmetric Dirichlet distribution, we can derive the Carnapian λ rules:

$$p(Q_{t+1}^q|E_t) = \frac{t_q + \lambda/2}{t + \lambda} = pr_\lambda(t_q, t). \tag{5}$$

A higher peak in the Dirichlet density $p(H_\theta)$ is reflected in a larger parameter λ.

Although the hypotheses in the above inferences can thus be replaced with direct links between observations, there are good reasons for keeping the hypotheses in. First, they express the chance mechanism that is supposed to underlie the observations. For example, if the observations concern coin tosses, we know that the mechanism underlying the observations concerns constant and independent chances. Second, the hypotheses enable us to express further knowledge of the chance mechanisms in a prior probability over them. In the example, a normal coin motivates a prior over these chances that is strongly peaked at $\frac{1}{2}$, while a coin from a conjurer's box may have little probability at $\frac{1}{2}$ and more probability at 0 and 1. In the prior we can thus express knowledge of the chance mechanism that is not incorporated in the statistical hypotheses themselves. It is not always a straightforward matter to incorporate such knowledge in a direct prediction rule.

3. A degenerate partition. The paper now directs attention to inductive inferences using two duplicate subpartitions, which differ only in the entirely theoretical property that they posit different mechanisms underlying the observations. It may seem pointless to use such a degenerate partition. However, the mechanisms underlying the observations can motivate different prior probabilities over these subpartitions. One reason for using duplicate partitions is thus that they facilitate the expression in the prior probability assignment of knowledge of underlying mechanisms. But because these priors react to the updating operations differently, the partitions can be distinguished by the observations after all, even while they consist of statistically identical hypotheses. Moreover, there are computational advantages to keeping the two subpartitions distinct. The upshot of this section is thus the same as the upshot of the previous one: theoretical notions, such as chance and mechanism, are useful for translating our knowledge into the prior probabilities that serve as input to a Bayesian statistical inference.

Let me start with the example on coin tosses. Imagine that we are undecided on whether the coin is from a conjurer's box or from an ordinary wallet. Now we are sure that the coin tosses are identical and independent trials, but in either case we are not sure what chance on tails the coin has. In sum, both these kinds of coins have an unknown constant chance θ on tails, say $q = 1$, but we have no hard restrictions otherwise. However, we do have some further knowledge of the mechanism underlying the observations that must somehow be incorporated in the statistical inference: if the coin is from the wallet, it is most probably fair, having a chance θ that is close to $\frac{1}{2}$, and if the coin is from the conjurer's box, it is most probably strongly biased, having a chance θ that is close to 0 or 1. On the other hand it may be a wallet coin with an unusual division of weight, corresponding with a chance away from half, and it may also be a coin from a rather cheap conjurer's box which does not show the intended deviant behaviour and has a chance quite close to half.

In order to incorporate this uncertain knowledge, we can employ an additional partition $\mathcal{G} = \{G_0, G_1\}$, referring to the normal and the magical coin respectively. Both hypotheses G_j cover exactly the same subpartition, $G_j = \{g_j\} \cdot \mathcal{H}$. They are only labelled differently. We can use the likelihoods θ for the hypotheses $g_0 \cdot H_\theta$

and $g_1 \cdot H_\theta$ alike. Note that in terms of these statistical hypotheses, the distinction between the magical coin and the normal coin is therefore not observable. For each hypothesis in the one subpartition, there is a hypothesis in the other subpartition that has exactly the same likelihoods for all the observations. The partition as a whole is in this sense degenerate.

There is a particular advantage, however, to employing the degenerate partition in the Bayesian scheme. We have separate control over the priors defined on the subpartitions on the normal and magical coin, $g_0 \cdot \mathcal{H}$ and $g_1 \cdot \mathcal{H}$ respectively. The further knowledge about the two kinds of coins, as described above, can motivate specific forms for the priors in both subpartitions. Specifically, we may decide to assign two different symmetric Dirichlet distributions, one that is sharply peaked at $\theta = \frac{1}{2}$ over $g_0 \cdot \mathcal{H}$, and one that is peaked at both $\theta = 0$ and $\theta = 1$ over $g_1 \cdot \mathcal{H}$. Such a chioce reflects the fact that we expect the normal coin to be fair, and the magical coin to be biased, while we are not sure of these expectations.

In the same way as in the foregoing, the prior over the degenerate partition leads to predictions over coin tosses. First, following the exposition of section 2, the two priors lead to two Carnapian rules, with $\lambda = 10$ for the peak at the centre and $\lambda = \frac{1}{4}$ for the two peaks at the sides. The priors are illustrated in figure 1. Furthermore, we are initially undecided between the two hypotheses on the origin of the coin, $p(G_0) = p(G_1)$. To arrive at predictions we can thus weigh the two Carnapian rules with the probabilities of the coin's origin, resulting in what Skyrms (1993) calls a hyper-Carnapian prediction rule. Again, the prediction rule can also be taken as an estimator of the parameter θ:

$$p(Q^q_{t+1}|E_t) = p(G_0|E_t)\, pr_{10}(t_q, t) \;+\; p(G_1|E_t)\, pr_{1/4}(t_q, t). \tag{6}$$

The idea is that the probabilities within the two subpartitions $g_0 \cdot \mathcal{H}$ and $g_1 \cdot \mathcal{H}$ are updated separately, and that the resulting values yielded by the Carnapian rules can function as the likelihoods in an update over the hypotheses G_0 and G_1.

Interestingly, even while the subpartitions associated with G_0 and G_1 consist of pairwise identical hypotheses, the differing priors over them cause different aggregated likelihoods of G_0 and G_1, namely the different Carnapian rules.

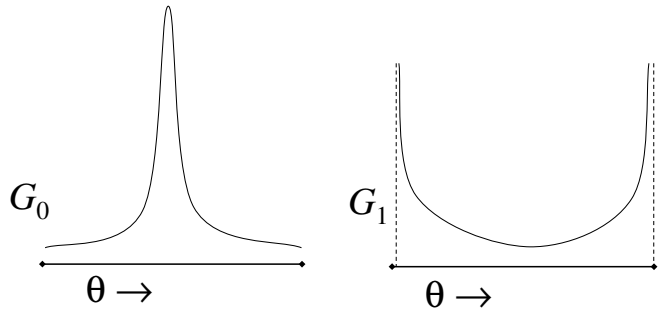

Figure 1: Two different priors over the two subpartitions of Bernoulli processes H_θ. The peak prior is associated with the normal coin, the valley prior with the magical coin. Both function are from the class of Dirichlet priors. For $\lambda = 2$ the prior distribution is uniform, for larger values of λ the peak gets higher, and for smaller values of $\lambda 2$ the valley gets deeper.

Therefore, if we update with a sequence of observations for which the relative frequency is quite close to 1, say $E_6 = 001000$, we will find that the updated probability of G_1 is larger than that of G_0. That is, the two subpartitions $g_j \cdot \mathcal{H}$ are observationally indistinguishable, but the different expectations over these partitions, as expressed in the different Dirichlet priors, make the partitions observationally distinct after all. In other words, as a side effect of updating over the separate subpartitions, the observations become relevant to the theoretical distinction between hypotheses G_0 and G_1.

This confirmation of theoretical hypotheses is less magical than it may seem. Indeed the distinction between the hypotheses G_0 and G_1 is theoretical in the sense that the hypotheses consist of the same subpartitions $\mathcal{H}$. But the hypotheses G_0 and G_1 do have different observational content: a magical coin is much less likely than the normal coin to yield an observed relative frequency of tails close to $\frac{1}{2}$. This content is exactly expressed in the differing priors over the subpartitions $g_0 \cdot \mathcal{H}$ and $g_1 \cdot \mathcal{H}$. The theoretical distinction between G_0 and G_1 simply facilitates the use of these differing priors over the two subpartitions. It can further be noted that the theoretical distinction helps in keeping the calculations manageable. It is possible to work with a single partition $\mathcal{H}$. The function that expresses the combined prior over this single partition is naturally the sum of the priors de-

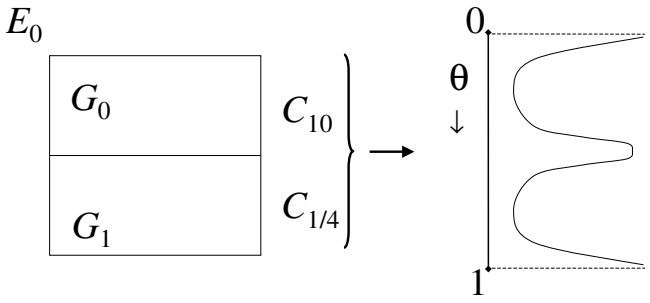

Figure 2: The hyper-Carnapian prediction rule can also be expressed in a single prior over one partition C. The prior is simply the sum of the two separate priors over the separate subpartitions.

fined over the above subpartitions, as expressed in figure 2, but it is much more convenient to update these terms separately. The resulting predictions cannot be equated with a single Carnapian rule, and it is not easy to find some other exchangeable direct prediction rule that captures them. The theoretical distinction thus helps to arrive at analytic results.

On the other hand, we can also turn this reasoning around. In the case of the above hyper-Carnapian rule, the reader may find that the representation of inductive predictions in terms of statistical hypotheses is entirely unhelpful, or even contrived. It may be much more natural to employ the hypotheses G_j with the Carnapian rules as likelihoods, and leave out the entire story on the underlying identical subpartitions $\mathcal{H}$, in which case there is no reason left for calling the distinction between the hypotheses G_j theoretical. However, there are some reasons for keeping the reference to the hypotheses H_θ on fixed chances in. First there are the reasons for employing the statistical hypotheses H_θ that I advenaced in section 2. Second, there are some problems in making sense of the Carnapian prediction rules in the role of statistical hypotheses. For one thing, none of the usual convergence theorems applies to them, because all such Carnapian hypotheses are continuously adapted to follow the observed relative frequencies.

Some more critical considerations concern the exact role of our knowledge

about underlying mechanisms, in this case knowledge about the possible origin of the coin. Note first that the duplicate partition is used to encode this specific knowledge into a prior probability assignment. More specifically, we motivate the shape of the prior using knowledge on the possible origin of the coin. However, I am not sure that we can speak of knowledge of the underlying mechanism here. In the example of the coin we are perhaps in that position, but in more interesting cases of scientific investigation the underlying mechanism can at best be a supposition on a few possible mechanisms. Following up on this, it is not in all cases clear how exactly such suppositions determine the shape of the prior probability assignment over the subpartitions. The example may suggest that this link is straightforward, but there are many cases in which the relation between supposition and prior is all but trivial.

Let me summarise the paper up to this point. I have shown how statistical hypotheses can be used in a Bayesian inference that yields predictions and parameter estimations. I briefly argued that the use of these hypotheses on chances serve a purpose: they express the chance mechanism that produces the observations, and they allow us to express further knowledge we may have on the mechanisms in a prior probability over them. After that I showed that the same holds for a degenerate partition, and in particular for the theoretical distinction employed in it. We included the theoretical distinction in the partition of hypotheses in order to allow ourselves to express and use all available knowledge on the chance mechanism that produces the observations. Indeed, I have not made clear the exact role of this knowledge, or these suppositions, in determining the priors. But I hope that I have sufficiently explained the use of degenerate partitions to consider its implications for the philosophy of science more generally.

4. The use of underdetermination. In this section, I want to tranfer the insights on the use of theoretical distinctions in statistical inference to scientific methodology more generally. The first claim deriving from this is that theoretical concepts offer a better grip on experimental testing in the two ways suggested above, namely by facilitating the expression of knowledge of underlying mechanisms in priors, and by carving up statistical inference in manageable parts.

In other words, the underdetermination created by theoretical notions serves a specific purpose in scientific method. The second claim follows from the observation that in cases like the above one, Bayesian statistics provides the reasons for choosing between hypotheses that are only theoretically distinct. Such choices are usually considered to involve abductive inference. We may therefore argue that the above cases provide a first sketch of a Bayesian model of abductive inference, countering claims by van Fraassen (1989) that inference to the best explanation is at variance with Bayesianism.

Let me first provide some more background for these claims. The problem of underdetermination is that science, if interpreted as a realist undertaking, is dramatically underdetermined by observation. At first sight it seems that much of the theoretical superstructure of scientific theories cannot be warranted by the observational substructure. The primary challenge, I submit, is to show that this apparent underdetermination is understandable in the light of the objective of science, where this objective, minimally, is to aim for the observational truth. The idea here is not to deny that there is underdetermination or to somehow resolve it, but to show that despite of the underdetermination we can make sense of scientific practice and method. Realists, however, take on the bigger challenge of showing that underdetermination can in some cases be avoided. One way they achieve this is by providing inference rules such as abduction, which enable us to choose between theoretical superstructures on the basis of explanatory considerations or other theoretical virtues. The underdetermination is in those cases resolved by employing these additional criteria.

Against this background I can be more precise about the idea of this section. In the following I start out with the first realist challenge and show that we can make sense of underdetermination. More specifically, I will show that underdetermined theoretical superstructures have a specific use in statistics. I thus accept that science is underdetermined, but I go on to suggest that it is possible to explain this fact by reference to the methodological use of underdetermination. Following up on this, I will argue that the use of theoretical superstructures looks a lot like abductive inference, as it provides the means to decide over models and hypotheses that are only theoretically distinct. But here the reason for the

question mark in the title will also become apparent, because I will claim that this does not come down to abduction simpliciter. While we can independently motivate the use of theoretical distinctions in partitions, the choice between theoretically distinct hypotheses within a partition is basically the result of our own expectations. I end by arguing that this amounts to a specific kind of abduction.

To substantiate the claim that underdetermination has a methodological use, recall from section 3 that the partition which employed purely theoretical distinctions offered a better grip on the statistical inferences from the observations. In the example, distinguishing the hypotheses G_0 and G_1 had a double use: it facilitated the expression of knowledge or suppositions on underlying mechanisms in priors, and carved up statistical inference in manageable parts. This already illustrates the methodological use of the degeneracy. In addition to this, consider a comparison between two statistical analyses, one in which the degenerate partition is used, and one that uses the single partition of section 2. For lack of any further theoretical story, we may apply entropy maximisation to arrive at a suitable prior over the hypotheses, or perhaps use a prior devised by Jeffreys (1939). However, simply because the prior over the degenerate partition is tailor-made to fit the most likely courses of events it will, if the coin is indeed from a wallet or a conjurer's box, converge to the correct predictions and the true parameter value more quickly. The statistical inference using the degenerate partition thus outperforms the single partition, and this is no surprise because the former employed more relevant knowledge as input. The point here is that the theoretical distinction and the accompanying story allowed us to use this knowledge. In other words, theoretical distinctions

On this point we should recall that, as a side effect of using theoretical distinctions, it looks as if these distinctions become observational. In the example, the updated probability of the hypothesis G_1 is larger than that of G_0 whenever the observed relative frequency is close to 0 or 1, and conversely when it is close to $\frac{1}{2}$, even while these two hypotheses consist of the very same statistical hypotheses. This is where the use of theoretical distinctions in statistical inference begins to look like abduction. Recall that an abductive inference enables us to choose between a number of observationally indistinguishable, and thus theoreti-

cal, alternatives on the basis of theoretical virtues, for example explanatory force. Now the suggestion here is that such a theoretical virtue is also presented by the fact that one of the two priors in the degenerate partition corresponds better to the observations. Recall that the observations have exactly the same impact on the separate hypotheses in each of the two subpartitions. The different impact is entirely due to the difference in the subjectively determined prior probability over these two subpartitions. We may therefore say that the observations reflect differently on the two subpartitions, because the observations interact differently with our theoretically motivated prior opinions.

5. Concluding remarks. To sum up, I have argued for two claims: there is a specific use for underdetermination in statistical inference, and in this use we encounter a peculiar kind of abductive inference. This abductive inference hinges on the interaction between the observations and our prior opinions. As for the use of underdetermination, I dare say that it generalises to science more generally. I believe that in many cases scientists employ theoretical superstructure to allow themselves better ways of using all available knowledge in testing and confirming their theories. But as indicated, I am not so sure about the claims on abduction. There is a sense in which the foregoing does present a Bayesian explication of abduction: the probability kinematics of the above presents an explanation of the fact that empirical scientists sometimes feel that they have reasons to prefer one theoretical hypothesis over another. However, to make that claim precise, I must argue that we can indeed call hypotheses such as G_j empirically equivalent. And as indicated before, this seems to conflict with the fact that the hypotheses do have observational content. The eventual resolution of this conflict can be found in a careful analysis of the nature of observations and the notion of empirical equivalence. But until that time, it seems that real abduction is only for aliens.

References

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

De Finetti, B. (1937) 'Foresight: its logical laws, its subjective sources' in *Studies in Subjective Probability*, eds. Kyburg, H. and Smokler, H. (1964), New York: John Wiley, pp. 97–158.

Earman, J. (1992) *Bayes or Bust*, Cambridge MA: MIT Press.

Howson, C. and Urbach, P. (1989) *Scientific Reasoning, The Bayesian Approach*, La Salle: Open Court.

Festa, R. (1993) *Optimum Inductive Methods*, Dordrecht: Kluwer.

Good, I. J. (1955) *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*, Cambridge (MA): MIT press.

Hintikka, J. (1970) 'Unknown Probabilities, Bayesianism, and De Finetti's Representation Theorem' in *Boston Studies in the Philosophy of Science*, Vol. VIII, eds. Buck, R. C. and Cohen, R. S., Dordrecht: Reidel.

Jeffreys, H. (1939) *Theory of Probability*, Oxford: Oxford University Press.

Skyrms, B. (1993) 'Analogy by Similarity in Hyper–Carnapian Inductive Logic', in J. Earman, A.I. Janis, G. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds*, Pittsburgh: University of Pittsburgh Press, pp. 273–282.

Van Fraassen, B. C. (1989) *Laws and Symmetry*, Oxford: Clarendon Press.