CrossMark

A Bayesian Solution to the Conflict of Narrowness and Precision in Direct Inference

Christian Wallmann[1]

Abstract The conflict of narrowness and precision in direct inference occurs if a body of evidence contains estimates for frequencies in a certain reference class and less precise estimates for frequencies in a narrower reference class. To develop a solution to this conflict, I draw on ideas developed by Paul Thorn and John Pollock. First, I argue that Kyburg and Teng's solution to the conflict of narrowness and precision leads to unreasonable direct inference probabilities. I then show that Thorn's recent solution to the conflict leads to unreasonable direct inference probabilities. Based on my analysis of Thorn's approach, I propose a natural distribution for a Bayesian analysis of the data directly obtained from studying members of the narrowest reference class.

Keywords Reference class problem · Direct inference · Statistical syllogism · Imprecise probabilities · Expected frequency · Specificity · Natural distribution

1 Introduction

In direct inference, the probability that a certain individual belongs to a target class is equated with the relative frequency of the target class in a suitable reference class (Venn 1888; Reichenbach 1949; Pollock 1990; Kyburg 1961). For instance, from the premises that (1) Roland is male Austrian and that (2) 27.3% of male Austrians smoke, it follows by direct inference that Roland smokes with probability 0.273. Knowledge of frequencies in narrower reference classes to which the individual belongs defeats the above direct inference. If we add the premises that (3) Roland is 32 years old and that (4) 35.7% of male Austrians aged 32 smoke, then it follows by direct inference that Roland smokes with probability 0.357. The single-case probability here is the rational degree of belief or the

✉ Christian Wallmann
c.wallmann-520@kent.ac.uk

[1] Department of Philosophy, University of Kent, Canterbury, Kent CT2 7NZ, UK

rational credence of an agent that entertains the relevant body of evidence. In what follows, I will call it *direct inference probability*. The direct inference probability expresses the agent's lack of knowledge and uncertainty regarding Roland's smoking status. It is not a physical probability, i.e., it does not express objective randomness in the world. For the sake of brevity, I often will not mention 'target class' or 'frequency' explicitly. For instance, instead of speaking of 'the estimate for the frequency of the target class in the narrower reference class', I will be speaking of the 'estimate for the narrower reference class'.

As pointed out by numerous authors, striving for narrow reference class and striving for precise estimates of frequencies in reference classes are competing aims (Venn 1888; Reichenbach 1949; Kyburg and Teng 2001). Narrower reference classes should be preferred in direct inference, because frequencies of the target class in narrower reference classes are more relevant to the direct inference probability. However, narrower reference classes have less members. Hence, it is harder to take a sufficiently large sample of individuals that belong to a narrow reference class. As a result statistical estimates for frequencies of the target class in a narrow reference class are often less precise. As the pioneer of reference class reasoning John Venn puts it

> [...]; but whilst cautioning us against appealing to too wide a class, it seems to suggest that we cannot go wrong in the opposite direction, that is in taking too narrow a class. And yet we do avoid any such extremes. John Smith is not only an Englishman; he may also be a native of such a part of England, be living in such a Presidency, and so on. An indefinite number of such additional characteristics might be brought out into notice, many of which at any rate have some bearing upon the question of vitality. Why do we reject any consideration of these narrower classes? We do reject them, but it is for what may be termed a practical rather than a theoretical reason. Now many of the attributes of any individual are so rare that to take them into account would be at variance with the fundamental assumption of our science, viz. that we properly concerned only with averages of large numbers (Venn 1888, p. 220).

Simply rejecting these narrower reference classes is a luxury we often cannot afford. In personalized medicine, for instance, the narrowest reference class for which reliable statistics can be compiled is often not narrow enough. Probabilistic relationships between the narrowest reference class for which reliable statistics can be compiled and the target class are often not strong enough to be useful for medical diagnosis and prediction (Manolio et al. 2009; Salari et al. 2012).

Hence, often a body of evidence contains estimates for frequencies in a certain reference class and less precise estimates for frequencies in a narrower reference class. In this case I say that a *conflict of narrowness and precision* obtains. To provide the rational degree of belief in face of a conflict of narrowness and precision amounts to *solving* it. What should our degree of belief that Roland smokes be if our body of evidence, for instance, contains in addition to (1) Roland is male Austrian and Roland is 32 years old and (2) 27.3% of male Austrians smoke, also that (3) the frequency of 32 year old male Austrians that smoke is between 20 and 50%?

To develop a new solution to the conflict of precision and narrowness, I draw on ideas developed by Paul Thorn and John Pollock. To determine direct inference probabilities, Thorn (2012, 2016) and Pollock (1990, 2011) consider arbitrary subsets of the broader reference class. I show that Thorn's approach leads to unreasonable direct inference probabilities and propose a remedy.

The paper is organised as follows. In Sect. 2, I present Kyburg and Teng's approach to direct inference. I argue that it leads to unreasonable direct inference probabilities. In Sect. 3, I present Thorn's and Pollock's approaches to direct inference. I then show that Thorn's solution to the conflict of narrowness and precision leads to unreasonable direct inference probabilities. I also analyse the causes for the failure of Thorn's solution. In Sect. 4, I propose a new Bayesian solution to the conflict of precision and narrowness that employs a natural prior.

2 Kyburg and Teng's Approach to Direct Inference

In this section, I present Kyburg and Teng's solution to the conflict of narrowness and precision. I argue that the resulting direct inference probabilities are often unreasonable. Instead of only considering the case of point-valued information, Kyburg and Teng also consider interval-valued information about frequencies in reference classes. Let T be a target class, $R, R', R_1, \ldots, R_5$ be reference classes, and c an individual. Suppose that $freq(T|R) \in [x, y]$ states that the relative frequency of individuals in R that are also T lies in the interval $[x, y]$. Finally, assume that $PROB(A)$ is the rational degree of belief that the proposition A is true. Following Kyburg and Teng, I call two intervals $[x, y]$ and $[u, v]$ *conflicting* if and only if neither one is a subset of the other, i.e., if and only if $[x, y] \not\subseteq [u, v]$ and $[u, v] \not\subseteq [x, y]$. Kyburg and Teng (2001) propose the following reference class rules.

> **Criterion of Precision** Suppose that $[x, y]$ and $[u, v]$ and $[x, y] \subseteq [u, v]$ are not conflicting and that $R' \subseteq R$. If $freq(T|R) = [x, y]$ and $freq(T|R') = [u, v]$, then $PROB(c \in T) = [x, y]$.

> **Criterion of Specificity** Suppose that $[x, y]$ and $[u, v]$ are conflicting and that $R' \subseteq R$. If $freq(T|R) = [x, y]$ and $freq(T|R') = [u, v]$, then $PROB(c \in T) = [u, v]$.

> **Combination of competing reference classes** Suppose that $R_1 \not\subseteq R_2$ and $R_2 \not\subseteq R_1$. If $freq(T|R_1) = [x, y]$ and $freq(T|R_2) = [u, v]$, then $PROB(c \in T) = [min\{x, u\}, max\{y, v\}]$.

The Criterion of Precision requires that in the case of non-conflicting intervals the more precise interval should be preferred. For instance, if $R' \subseteq R$, $freq(T|R) = [0.3, 0.5]$, $freq(T|R') \in [0.25, 0.7]$, and $c \in R'$, then $PROB(c \in T) = [0.3, 0.5]$. The Criterion of Specificity states that in the case of two conflicting intervals the narrower reference class should be preferred. For instance, if $R' \subseteq R$, $freq(T|R) = [0.4, 0.6]$, $freq(T|R') \in [0.5, 0.8]$, and $c \in R'$, then $PROB(c \in T) = [0.5, 0.8]$. The combination rule demands that $PROB(c \in T)$ is located in the convex hull of the intervals for the competing reference classes. For instance, if $freq(T|R_1) = [0.3, 0.5]$, $freq(T|R_2) = [0.4, 0.7]$, $c \in R_1$ and $c \in R_2$, then $PROB(c \in T) = [0.3, 0.7]$.

If a body of evidence contains information for many different reference classes, the criteria of specificity and precision may interact. For such complex direct inference scenarios Kyburg and Teng (2001) suggest to *sharpen* the body of evidence before drawing any direct inference. They propose the following procedure to sharpen a body of evidence. In a first step, they apply the Criterion of Specificity to the whole body of evidence. This step rules out all intervals that are conflicting with the interval for the most specific reference class. In a second step, they apply the Criterion of Specificity to the second most specific reference class in the remaining set. They iterate this procedure until for any two

remaining reference classes $R' \subseteq R$ the intervals are not conflicting. Application of the Criterion of Precision yields the tightest of these intervals. Finally, they apply the combination rule to the remaining competing reference classes. The following example illustrates this procedure. Suppose that c belongs to the reference classes R_1, R_2, R_3, R_4, R_5. Our body of evidence contains the following information about frequencies in those reference classes.

$$freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4 \wedge R_5) = [0.2, 0.8]$$
$$freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4) = [0.19, 0.6]$$
$$freq(T|R_1 \wedge R_2 \wedge R_3) = [0.4, 0.8]$$
$$freq(T|R_1 \wedge R_2) = [0.2, 0.3]$$
$$freq(T|R_1) = [0.21, 0.22]$$

First, $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4 \wedge R_5) = [0.2, 0.8]$ rules out $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4) = [0.19, 0.6]$ by specificity. Second, $freq(T|R_1 \wedge R_2 \wedge R_3) = [0.4, 0.8]$ rules out $freq(T|R_1 \wedge R_2) = [0.2, 0.3]$ and $freq(T|R_1) = [0.21, 0.22]$ by specificity in the remaining set. The remaining set $\{freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4 \wedge R_5) = [0.2, 0.8], freq(T|R_1 \wedge R_2 \wedge R_3) = [0.4, 0.8]\}$ consists of two non-conflicting intervals. $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4 \wedge R_5) = [0.2, 0.8]$ is ruled out by precision. Hence, $PROB(c \in T) = [0.4, 0.8]$. I now argue that Kyburg and Teng's approach leads to unreasonable direct inference probabilities. First, their overemphasis on the fact that two intervals conflict leads to unstable inference. Minimal changes in reference class probabilities may lead to huge changes in the direct inference probability. Consider, for instance, the reference class information $freq(T|R') = [0.2, 0.6]$ and $freq(T|R) = 0.2$. The Criterion of Precision leads to $PROB(c \in T) = 0.2$. Consider now the small change from $freq(T|R) = 0.2$ to $freq(T|R) = 0.19$. The Criterion of Specificity yields $PROB(c \in T) = [0.2, 0.6]$ rather than $PROB(c \in T) = 0.2$. This instability carries over to more complex bodies of evidence. If, for instance, in the above example $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4) = [0.19, 0.6]$ is replaced by $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4) = [0.2, 0.6]$, then $PROB(c \in T) = [0.21, 0.22]$.[1]

Second, Kyburg and Teng treat all values in the interval of all possible frequencies equally. However, in practice, these values are estimated from a sample. Values within the center of those intervals maximize the likelihood of the observed sample. To determine the direct inference probability, positions of points within the interval should therefore be taken into account (see also Sect. 4.2). In the above example, if obtained by a sample, the maximum likelihood estimate for $freq(T|R')$ lies in most cases in the center of the interval [0.2, 0.6]. Hence, the sample of the narrower reference class provides evidence against the value $freq(T|R') = 0.2$. Hence, it is unreasonable to extrapolate the frequency 0.2 from the broader reference class to the narrower reference class. As a consequence, Kyburg and Teng's approach may lead to irrational decisions (Stone 1987, pp. 253–254). Stone considers the case in which the Ace Urn Company places balls ordered into urns. We get the following information.

1. The Ace Urn Company orders a proportion of .51 red balls;
2. The proportion of red balls ordered by the Taiwan Division is somewhere in [.01, .52].

---

[1] In this case, $freq(T|R_1 \wedge R_2 \wedge R_3 \wedge R_4) = [0.2, 0.6]$ remains in the body of evidence after carrying out step 1. In a second step, $freq(T|R_1 \wedge R_2 \wedge R_3) = [0.4, 0.8]$ is removed from the body of evidence. Therefore, $freq(T|R_1 \wedge R_2) = [0.2, 0.3]$ and $freq(T|R_1) = [0.21, 0.22]$ remain in the body of evidence. Applying precision, we obtain $PROB(c \in T) = [0.21, 0.22]$.

An agent draws a ball from an urn which is labeled by "Ace Urn Company-Made in Taiwan". What is the probability that the agent will draw a red ball? Kyburg and Teng's approach leads to the direct inference probability 0.51. Suppose a brilliant chef offers us a very delicious meal, if either (1) we guess the correct color of the ball or (2) we roll a fair die and it comes up 6. We have the choice. I agree with Stone that we should choose the first option and guess "black". Kyburg and Teng, however, recommend "red".

3 Subset Approaches

For determining direct inference probabilities, Pollock (2011) and Thorn (2016) consider arbitrary subsets of the broader reference class. In this section, I present Thorn's solution to the conflict of narrowness and precision. I show that and why it leads to unreasonable direct inference probabilities. Finally, I argue that Pollock's approach to direct inference encounters the very same problems.

3.1 Thorn's Epistemic Utility Argument

Thorn (2016) gives the following epistemic utility argument for equating the direct inference probability with the expected frequency in the narrowest reference class. Denote by 'S_c' the conjunction of all properties that the individual c is known to have. Suppose that we assign to *all* individuals $d_1, \ldots, d_n$ with the properties S_c the same direct inference probability $Prob(d_i \in T) = v_i = v$. Thorn calls such a policy *principled*. Let $V(d_i \in T) \in \{0, 1\}$ be the truth value of the proposition $d_i \in T$. Assume that we measure epistemic inaccuracy by the average squared deviation of the predicted values from the true values

$$S = \frac{1}{n} \sum_{i=1}^{n} (v - V(d_i \in T))^2.$$

Then S is minimised if and only if $v = freq(T|S_c)$. Hence, setting the direct inference probability to the frequency in the narrowest reference class maximises epistemic accuracy in the class S_c. Thorn (2016) shows that this result holds for a much more general class of accuracy measures (so-called proper scoring rules).

Often, however, $freq(T|S_c)$ will be unknown. In these cases, the expected value of $freq(T|S_c)$ maximises expected accuracy in the class S_c (Thorn 2016). Thorn calls the expected value of $freq(T|S_c)$ *expected frequency*. Following Thorn, we denote it by '$E[freq(T|S_c)]$'. For this reason he entertains the following direct inference rule:

> If R' is the narrowest reference class the individual c is known to belong to and $E[freq(T|R')] = r$, then $PROB(c \in T) = r$.[2]

This concludes the presentation of Thorn's argument. Thorn (2016) himself identifies and discusses four non-trivial assumptions in his argument. First, the expected frequency in the narrowest reference class minimises expected inaccuracy only if accuracy is measured by proper scoring rules. If epistemic inaccuracy is measured by the absolute deviation $S' = \frac{1}{n} \sum_{i=1}^{n} |v - V(d_i \in T)|$, then the expected frequency in the narrowest reference class

---

[2] Note that Thorn's argument does not justify the following version of the principle of the narrowest reference class: If we know that $Dc, Bc, Cc, freq(T|B) = r, freq(T|B \wedge C) = s$, then $PROB(c \in T) = s$. The additional assumption $E[freq(A|B \wedge C \wedge D)] = s$ is needed.

does not minimise S'. Second, the expected frequency in the narrowest reference class minimises expected inaccuracy only if we make predictions about all individuals with S_c. Third, expected accuracy is relative to the distribution employed to calculate the expected accuracy. Maximising expected accuracy is only a legitimate aim if the distribution is empirically accurate, i.e., if it matches the relative frequencies in the world. Fourth, even if the distribution employed to calculate the expected frequency is accurate, one needs to care about expected accuracy.

Thorn's assumptions are by no means uncontroversial. Below, I discuss ways in which they may fail. For Thorn's second assumption, two alternatives seem to be at least equally plausible. We could aim to make the best decision for the single-case $c \in T$ or we may require maximum accuracy in the class of all individuals for which predictions are being made. In both cases setting $PROB(c \in T) = freq(T|S_c)$ does not maximise accuracy. Regarding the third assumption I show in Sect. 4.1 that Thorn's distribution to determine the expected value and expected accuracy is inaccurate. Finally, clearly, expected accuracy is of main importance in the long-run—but should it be of importance in the short-run? Opinions are divided here. On the one hand, Pollock claims that many single-cases add up to a long-run. For this reason, they should not receive special treatment.

> People sometimes protest at this point that they are not interested in the general case. They are concerned with some inference they are only going to make once. They want to know why they should reason this way in the single case. But all cases are single cases. If you reason in this way in single cases, you will tend to get them right (Pollock 2011, p. 32).

On the other hand, Williamson does not require that direct inference probabilities maximise expected accuracy. His direct inference probabilities should minimise worst-case expected loss (Williamson 2013). Minimising worst-case loss is a cautious strategy.[3]

A detailed discussion of Thorn's assumptions goes beyond the scope of the present paper. In what follows I mainly show that Thorn's third assumption is violated. His distribution employed to calculate expected accuracy is inaccurate. Hence, maximising accuracy relative to Thorn's distribution is the wrong thing to do.

3.2 Thorn's Solution to the Conflict of Narrowness and Precision

Consider now the conflict of narrowness and precision. Assume that c belongs to two reference classes $R' \subseteq R$. Suppose that 1) we have precise-valued information $freq(T|R) = x$ for the frequency in the broader reference class and that 2) we have imprecise-valued information $freq(T|R') = v_1 \vee \ldots \vee freq(T|R') = v_n$ for the frequency in the narrower reference class. According to Thorn's approach (see Sect. 3.1), $PROB(c \in T) =$

---

[3] In the following example, minimising worst-case loss and maximising accuracy lead to different direct inference probabilities. Suppose, for instance, that we only know that c belongs to the reference class E. Suppose further that a statistical trial yields the maximum likelihood estimate m for $freq(T|E)$, i.e., setting $m = freq(T|E)$ maximises the probability of the observed outcome of the trial. Williamson recommends to calibrate $PROB(c \in T)$ only to a certain extent with the maximum likelihood estimate m. Suppose that $[m - a, m + a]$ is a $x\%$-confidence interval for $freq(T|E)$, where x is the confidence level at which the agent grants that $freq(T|E) \in [m - a, m + a]$. If the agent grants that $freq(A|E) \in [m - a, m + a]$, Williamson's calibration norm locates $PROB(c \in T)$ in the interval $[m - a, m + a]$. Williamson's equivocation norm selects the most cautious value in $[m - a, m + a]$, i.e., whichever is the closest to 0.5, for $PROB(c \in T)$. If confidence intervals are wide and estimates are less precise, then Williamson's direct inferences may differ considerably from the maximum likelihood estimate m. For instance, if $m = 0.25$ and a 95%-confidence interval is [0.1, 0.4], then Williamson's approach leads to $PROB(c \in T) = 0.4$.

$E[freq(T|R')]$. To determine the direct inference probability we therefore need to calculate the expected frequency $E[freq(T|R')]$. Thorn proposes the following method to do this.

In what follows we set $V = \{v_1, \ldots, v_n\}$ so that $freq(T|R') = v_1 \vee \ldots \vee freq(T|R') = v_n$ becomes $freq(T|R') \in V$. According to Thorn's Method 1, $E[freq(T|R')]$ is a weighted average of the values in V (Thorn 2016). The weights are the probabilities that $freq(T|R')$ has value v_i, i.e., $PROB(freq(T|R') = v_i)$. Thorn proposes to determine these probabilities by drawing direct inferences for the reference class R'. Since this strategy treats reference classes as individuals and subsumes them under other reference classes, I call the resulting direct inferences *meta direct inferences*.

According to Thorn, the most appropriate (meta) reference class for R' in this case is the set of all subsets of R that have the same size as R' and whose relative frequency of the target class is among the values in V. Hence, $PROB(freq(T|R') = v_i) = freq(\{S : freq(T|S) = v_i\}|\{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V\}) := p_i$ (Thorn 2016). Note that $p_i = \frac{z_i}{\sum_{i=1}^{n} z_i}$, where

$$z_i = freq(\{S : freq(T|S) = v_i\}|\{S : S \subseteq R \wedge |S| = |R'|\}). \tag{1}$$

Means to calculating the z_i are well-known in finite combinatorics. In fact, the z_i are hypergeometrically distributed (see "Appendix" section Equation (5)). We will come back to this in Sect. 3.4. Thorn's reasoning can be summarized by the following two steps.

Step 1: Meta direct inference to determine the weights
Premise 1:
$p_i = freq(\{S : freq(T|S) = v_i\}|\{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V)\})$
Premise 2: $R' \in \{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V\}$
Conclusion: $PROB(freq(T|R') = v_i) = p_i$

Step 2: Build weighted average of possible values for the narrower reference class

$$E[freq(T|R')] = \sum_{i=1}^{n} v_i \times PROB(freq(T|R') = v_i) = \sum_{i=1}^{n} v_i \times p_i. \tag{2}$$

The method generalizes then to the case $|R'| \in W$ where W is an arbitrary set (see Thorn 2016, Theorem 5).

To calculate $PROB(freq(T|R') = v_i)$ by the above meta direct inference is initially plausible. If the frequency of smokers in Austria is 30% and Roland is Austrian, then the probability that Roland smokes is 0.3. Analogously, if the frequency of subsets S of R such that $freq(T|S) = v_i$ is p_i and if R' is a subset of R, then the probability that $freq(T|R') = v_i$ is p_i. As long as no better (meta) reference class for R' can be found, conclusions drawn by Thorn's approach remain plausible. In Sect. 4.1, I show that a better reference class can be found.

Thorn illustrates Method 1 by means of the following example:

Example 1 Suppose Bill is a member of Company B. Company B has 100 members and 25 members are NCOs. Suppose also that Bill is a member of the command unit of Company B. The command unit has 10 members. Either 20% of the command unit are NCOs or 30% are NCOs. What is the probability that Bill is an NCO?

The example can be formalised this way: $|R| = 100$, $freq(T|R) = 0.25$, $|R'| = 10$ and $freq(T|R') = 0.2 \vee freq(T|R') = 0.3$. Following his Method 1, Thorn derives that the probability that Bill is NCO $PROB(c \in T)$ is 0.2485.[4]

As Thorn correctly claims, this result does not depend much on the size of Company B. In addition, as we will shortly see, if $|R'|$ is sufficiently large, $E[freq(T|R')]$ does not depend much on the size of the command unit either.

3.3 Counter-Examples to Thorn's Method

In this section, I give some numerical examples in which Thorn's Method 1 leads to unreasonable direct inference probabilities.[5] These examples show that the weights p_i are too extreme to serve as basis for direct inference.

First, Thorn's Method 1 is not monotonic in the following sense. Suppose that $freq(T|R) = 0.25$. If we compare $freq(T|R') = s \vee freq(T|R') = t$ with $freq(T|R') = s \vee freq(T|R') = t'$, where $t < t'$, the later may result in a lower value for $E[freq(T|R')]$. In Thorn's example, with $freq(T|R') = 0.2 \vee freq(T|R') = 0.4$, we obtain $E[freq(T|R')] = 0.266$, with $freq(T|R') = 0.2 \vee freq(T|R') = 0.5$, we obtain $E[freq(T|R')] = 0.246$, and finally with $freq(T|R') = 0.2 \vee freq(T|R') = 0.6$, we obtain $E[freq(T|R')] = 0.216$.

Second, assume that the possible frequencies for the narrower reference class span an interval that contains the frequency of the broader reference class. In this case Method 1, assigns most of the weight to the frequency closest to the frequency of the broader reference class. If $freq(T|R') = 0.2 \vee freq(T|R') = 0.7$, we obtain $E[freq(T|R')] = 0.203$. The weights are $w_1 = \frac{32589}{32798}$ and $w_2 = \frac{209}{32798}$. Hence, $\frac{w_1}{w_2} = 155.9282297$, i.e., w_1 is almost 156 times as high as w_2.[6]

Third, assume that the possible values for the narrower reference class span an interval that contains the frequency of the broader reference class, that $|R|$ and $|R'|$ are sufficiently high, and that $freq(T|R)$ is a possible value for $freq(T|R')$.[7] In this case the information about the narrower reference class is ignored by Method 1. In other words: Thorn's approach leads to Kyburg's Criterion of Precision (see Sect. 2). We have seen that the Criterion of Precision is flawed. For instance, if $|R| = 1000$ and $|R'| = 100$, and $freq(T|R') = 0.2 \vee freq(T|R') = 0.21 \dots \vee freq(T|R') = 0.39 \vee freq(T|R') = 0.4$, then $E[freq(T|R')] = 0.257$. This value will not change if we replace the endpoint of the interval by any value in the set $\{0.41, 0.42, \dots, 1\}$. Furthermore, this value will not change much if we vary $|R'|$. If $|R'| \in [100, 900]$, then the minimum of the corresponding expected frequencies is 0.249 and the maximum is 0.259. It is therefore reasonable to assume that $E[freq(T|R')] \in [0.249, 0.259]$.

Fourth, and related, consider the case in which the possible values for the narrower reference class span an interval that does not contain the value of the broader reference

---

[4] Step 1: $freq(\{S : freq(T|S) = 0.2\}|\{S : S \subseteq R \wedge |S| = 10 \wedge (freq(T|S) = 0.2 \vee freq(T|S) = 0.3)\}) = 0.515$ and $freq(\{S : freq(T|S) = 0.3\}|\{S :: S \subseteq R \wedge |S| = 10 \wedge (freq(T|S) = 0.2 \vee freq(T|S) = 0.3)\}) = 0.485$.
Step 2: $E[freq(T|R')] = 0.2 \times 0.515 + 0.3 \times 0.485 = 02485$.

[5] The numerical values are calculated by using the Matlab-code presented in the "Appendix" section.

[6] One may think that this fact does change if we move to a higher number for $|R'|$. Quite the opposite is the case: If $|R| = 1000$ and $|R'| = 100$, and $freq(T|R') = 0.2 \vee freq(T|R') = 0.4$, then $E[freq(T|R')] = 0.200$.

[7] Or if the closest values to $freq(T|R)$ are symmetrical around $freq(T|R)$.

class. In this case, the closest value to $freq(T|R)$ gets most weight. In many cases it gets almost all weight. For instance, if $freq(T|R') = 0.5 \vee freq(T|R') = 0.7$, then $E[freq(T|R')] = 0.507$.[8]

Of course, many of such examples can be given, but I assume that those given here are sufficient to show that Thorn's Method 1 is not correct.

3.4 Diagnosis

The main reason for the fact that Thorn's Method 1 leads to unreasonable direct inference probabilities is this: to build the needed expected frequency, Thorn relies on relative frequencies in arbitrary subsets of the broader reference class. Since, as I show in this section, the variation of these relative frequencies is very small, the weights p_i are extreme.
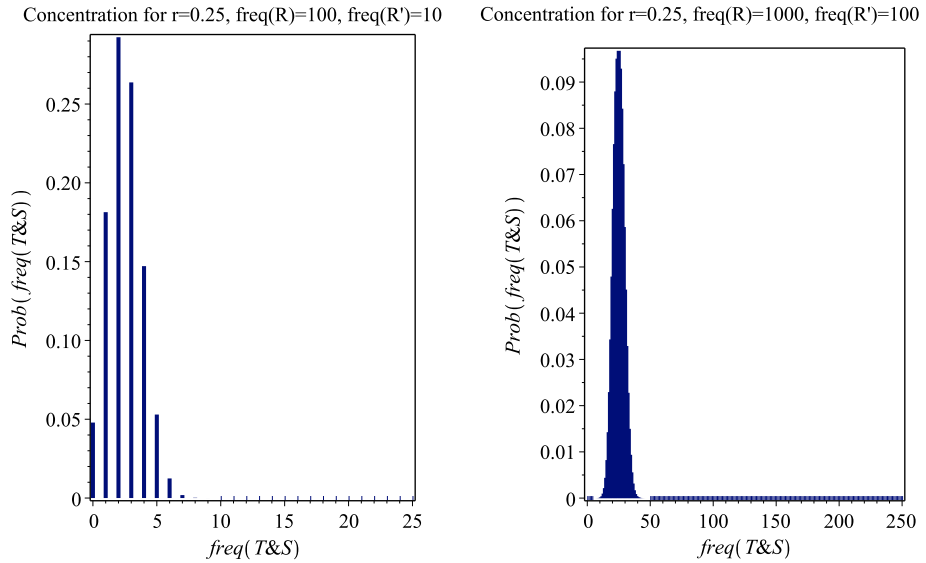
Suppose that $freq(T|R) = r$, then relative frequencies $freq(T|S)$ in subsets $S \subseteq R$ cluster around r (see "Appendix" section). Surprisingly, however, these relative frequencies are quite uniform. If the sizes of the sets R and R' are sufficiently large, then for almost all subsets $S \subseteq R$, $freq(T|S)$ is very close to r. In other words: The variance of the distribution of $freq(T|S)$ around its expected value r is small. Figure 1 illustrates this "concentration" or "peaking" property. For a more precise statement see Theorem 1 in the "Appendix" section. Moreover, if the sizes of the sets R and R' are tending to infinity, then for all $\epsilon > 0$ $freq(\{S : S \subseteq R \wedge |S| = |R'| \wedge |r - freq(T|S) > \epsilon|\})$ is tending to 0. I.e., for almost all subsets $S \subseteq R$ it is the case that $|freq(T|S) - freq(T|R)|$ is smaller than every fixed number (think of epsilon as, for instance, $\frac{1}{1000000000}$).[9]

In Example 1, for instance, the variance of the distribution of $freq(T|S)$ is approximately 0.017. It follows that for 95% of sets $S \in \{S : S \subseteq R \wedge |S| = |R'|\}$ it holds that $freq(T|S) \in [0, 0.5]$. Although this is a reasonable spread around 0.25, it still leads to problem one, two and four for Thorn's Method 1 as discussed in Sect. 3.3. Compared to values that are smaller than 0.5, values for $freq(T|R')$ that are higher than 0.5 get almost no weight. Worse still, since the variance of the relevant distribution is tending to zero as the sizes of R and R' become larger, this tendency magnifies. This "peaking" around the expected value is responsible for the third problem for Thorn's Method 1 as discussed in Sect. 3.3. For instance, if $|R| = 1000$ and $|R'| = 100$ in Thorn's example, then the variance is approximately 0.002. It follows that for 95% of sets $S \in \{S : S \subseteq R \wedge |S| = |R'|\}$ it holds that $freq(T|S) \in [0.17, 0.33]$.

I conclude that to determine the probabilities p_i, Thorn uses a distribution with very low variance. The resulting direct inference probabilities are therefore extreme. To improve Thorn's approach, there are two (not mutually exclusive) options. First, one may consider special subsets of the broader reference class. The resulting distribution may then have sufficiently high variance. In Sect. 4.1, I propose such a special class of subsets. Second, to obtain more balanced probabilities, one may combine the distribution obtained from considering subsets of the broader reference class with a second distribution. Indeed, that the weights p_i are solely determined by the distribution of subsets of the broader reference class is a second cause for the failure of Thorn's Method 1. Thorn ignores the fact that in most cases there has been evidence that establishes which values are epistemically possible for the narrower reference class. There has to be data or other evidence for the fact that, for

---

[8] This fact does not change if we move to higher numbers for $|R|$. Quite the opposite is the case: If $|R| = 1000$ and $|R'| = 100$, and $freq(T|R') = 0.3 \vee freq(T|R') = 0.4$, then $E[freq(T|R')] = 0.300$.

[9] The reason for this is that k occurs in the denominator of the variance of the distribution of $freq(T|S)$ (see Theorem 1 in the "Appendix" section).

Concentration for r=0.25, freq(R)=100, freq(R')=10   Concentration for r=0.25, freq(R)=1000, freq(R')=100

Fig. 1 Plot of $f : \{0, \ldots, |R|\} \to [0,1]$; $f(s) := PROB(X_{Freq} = s)$

instance, $freq(T|R') = 0.5 \vee freq(T|R') = 0.7$. If this evidence is quite good, then accurate weights should be less sensitive to the distribution of subsets of the broader reference class. In Sect. 4.2, I propose a Bayesian way to combine the probabilities obtained from the distribution of subsets of the broader reference class with probabilities obtained form data on the narrower reference class.

3.5 The Case of No Information Concerning the Narrower Reference Class

Suppose that no information about the frequency in the narrower reference class is available. To determine the direct inference probability in this case, Thorn applies Method 1 to the logical truth $freq(T|R') = \frac{0}{|R'|} \vee freq(T|R') = \frac{1}{|R'|} \vee \ldots \vee freq(T|R') = \frac{|R'|}{|R'|}$. Suppose that $freq(T|R) = r$ and $R' \subseteq R$. Then:

$$E[freq(T|R')] = \sum_{i=0}^{|R'|} \frac{i}{|R'|} PROB(freq(T|R') = \frac{i}{|R'|}).$$

To determine $PROB(freq(T|R') = \frac{i}{|R'|})$, Thorn applies Step 1 of Method 1. Method 2 then yields the following result (Thorn 2016),

If $freq(T|R) = r$ and $R' \subseteq R$, then $E[freq(T|R')] = r$.

I agree that the conclusion $E[freq(T|R')] = r$ is reasonable in such cases.[10] However, I think that the reasoning leading to this conclusion is faulty. To determine $E[freq(T|R')] = r$, the set of all subsets of the broader reference class is the wrong (meta)

---

[10] This is a consistency condition for direct inference: If the direct inference from $R' \subseteq R$, $c \in R'$ and $freq(T|R) = r$, to $PROB(c \in T) = r$ should go through and if $PROB(c \in T) = E[freq(T|R')]$, then $E[freq(T|R')] = r$.

reference class for R'. In Sect. 4.1, I will propose a more suitable (meta) reference class. However, expected values of two different distributions may agree. By a lucky coincidence Thorn's approach yields reasonable direct inference probabilities in this case. In other cases there is no such luck. Thorn's approach yields the wrong direct inference probabilities (see Sect. 3.3).

3.6 Pollock's Approach to Direct Inference

Like Thorn, Pollock motivates his theory of direct inference by considering the distribution of arbitrary subsets in sets.

> Suppose we have a set of 10,000,000 objects. I announce that I am going to select a subset, and ask you how many members it will have. Most people will protest that there is no way to answer this question. It could have any number of members from 0 to 10,000,000. However, if you answer, Approximately 5,000,000, you will almost certainly be right. This is because, although there are subsets of all sizes from 0 to 10,000,000, there are many more subsets whose sizes are approximately 5,000,000 than there are of any other size. In fact, 99% of the subsets have cardinalities differing from 5,000,000 by less than .08%. (Pollock 2011, p. 329)

Pollock (2011) equates the direct inference probability with what he calls the *expectable value* of the narrowest reference class. He calculates the expectable value within his theory of probable probabilities. These probabilities extrapolate the combinatorial probabilities for finite sets employed by Thorn to infinite sets. The expectable value only exists if the variance of the distribution of arbitrary subsets tends to zero as the size of the broader reference class tends to infinity. In Sect. 4.1, I show that the reasoning underlying Pollock's and Thorn's approach to direct inference is faulty. Considering arbitrary subsets of sets is not appropriate for determining direct inference probabilities. Consequently, although they may be accurate in some cases, in general, one cannot trust in the correctness of Pollock's direct inference probabilities.

4 Remedy: Natural Distributions

In this section, I propose a new Bayesian solution to the conflict of narrowness and precision. I argue that the meta direct inference to determine the weights (Step 1 in Thorn's Method 1) can be defeated. The set of all subclasses of the broader reference class R that people *actually use* in direct inference is a more suitable (meta) reference class for R' than the set of arbitrary subsets $POW(R, |R'|)$. The probabilities obtained by this (meta) reference class yield a natural prior distribution for my Bayesian approach.

4.1 Reference Classes are Exceptional Subsets

Thorn's Step 1 in Method 1 is based on meta direct inference. To draw the relevant direct inference, he subsumes R' under the reference class of all subsets of the broader reference class R. As I believe in the cogency of direct inference, in order to refute Thorn's Method 1, the meta direct inference has to be defeated by a narrower or a competing reference class for R'. Indeed, the set of all reference classes people actually use in direct inference is such a narrower reference class. Let $RefA(T)$ be the set of all reference classes with respect to the

target class T that are actually used in direct inference and let $p_i^* = freq(\{S : freq(T|S) = v_i\}|\{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V \wedge S \in RefA(T)\})$.

Subclass defeat for the meta direct inference to determine the weights
Premise 1: $p_i = freq(\{S : freq(T|S) = v_i\}|\{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V\})$
Premise 2: $R' \in \{S : S \subseteq R \wedge |S| = |R'| \wedge freq(T|S) \in V\}$
Premise 3: $R' \in RefA(T)$
Premise 4: $p_i \neq p_i^*$
Conclusion: $PROB(freq(T|R') = v_i) \neq p_i$

Premise 4 is reasonable. The fact that for almost all subsets S of the broader reference class R, $freq(T|S)$ is close to $freq(T|R)$ is difficult to reconcile with experience (see also Wallmann and Williamson 2017). In practice, we often find subsets that contain a rather different relative frequency of the target class than the original set. For instance, smoking rates in the United States vary strongly with gender, age, education, poverty status and many more. But according to the distribution of frequencies of all subsets, such variations are almost impossible (see Theorem 1 in the "Appendix" section). In other words: Thorn's probabilities *PROB* for frequencies in narrower reference classes do not match our observed frequencies for frequencies in narrower reference classes, i.e., they are inaccurate. But why is this the case?

In direct inference we consider certain classes of individuals, because we believe that they are causally related to the target class. Now, our past success in detecting causally relevant classes and the fact that almost all subclasses of reference classes are not causally relevant to the target class, suggest that we are quite successful in detecting "exceptional" subsets. Since they causally interact with the target class, these exceptional classes tend to be difference makers, i.e., the target class and the reference class tend to be probabilistically dependent. Therefore, frequencies within sub-reference classes that we actually use in direct inference do not cluster around a single value. They cluster around multiple values. Hence, Premise 4 is plausible: The variance among frequencies in sub-reference classes that we actually use in direct inference is higher than in subsets in general. Therefore, the new weights p_i^* are more balanced than the p_i.

We call distributions that describe how frequencies in sub-reference classes which we actually use in direct inference are distributed *natural distributions* (for the concept of natural distributions in a different context see Paris et al. 2000). We should use natural distributions in direct inference because they yield the best long-run epistemic consequences in the intended class: The class of all direct inferences that we actually draw. In absence of further knowledge, the natural distribution maximises expected accuracy in the class of all direct inferences that we actually draw (see Sect. 3.1).

Granted that the natural distribution is most suitable for direct inference. This fact is of little help for drawing direct inferences in practice, if there is no way to determine the natural distribution. How can we find out about the natural distribution? Paris et al. (2000) discuss two ways to estimate natural distributions in general. First, by empirical experimentation. We could (1) draw a sample of all direct inferences actually drawn such as $freq(T|R) = x$, (2) study direct inferences in which subclasses $S \subset R$ were employed, and finally (3) consider the frequencies $freq(T|S)$ for such $S's$. The distribution of these frequencies is then an estimate for the natural distribution. Second, we may propose some reasonable properties that natural distributions are supposed to have. For instance, we may assume the default independence $E[freq(T|R')] = freq(T|R) = x$.

A detailed discussion of how to find the natural distribution is beyond the scope of the present paper, but I think I have said enough to make three crucial points needed here. First, in building expected frequencies of the narrowest reference classes sub-reference classes that we actually use in direct inference (rather than arbitrary subclasses of the broader reference class) should be considered. Second, this natural distribution differs from Thorn's distribution. Third, it is possible, at least in principle, to determine the natural distribution.

4.2 A Bayesian Solution to the Conflict of Narrowness and Precision

On the one hand, in the empirical sciences, frequencies $freq(T|R')$ within reference classes are in most cases estimated by a suitable statistical procedure from a sample of members of R'. For instance, suppose that an observed random sample[11] (with replacement) of 16 R'-individuals contains 8 T-individuals. The likelihood for such a sample is $freq(T|R')^8 \cdot (1 - freq(T|R'))^8$. The likelihood is maximised for $freq(T|R') = 0.5$. The likelihood is much smaller if, for instance, $freq(T|R') = 0.1$. On the other hand, it should not be ignored that R' belongs to the set of all sub-reference classes of R people actually use. Thus, I reformulate the conflict of narrowness and precision this way.

> **Conflict of narrowness and precision reformulated** How should the following information be aggregated to get an estimate for the value of $freq(T|R')$? 1) Frequencies in a sample of the narrower reference class and 2) probabilities obtained from the natural distribution of sub-reference classes of R.

I propose to assign more weight to estimates based on data than to estimates based on the fact that the narrower reference class is a sub-reference class of the broader reference class. Especially, if the probability estimates based on data directly about the narrower reference and the probability estimate for the broader reference class have almost the same precision, this is reasonable.

> **Data over Expectation Principle** When determining an estimate for $freq(T|R')$, frequencies of the target class in narrower reference classes based on data should get more weight than probabilities derived from the fact that the narrower reference class is a sub-reference class of the broader reference class.

To satisfy the Data over Expectation Principle, I propose to use the natural distribution of sub-reference classes in the broader reference class as prior distribution in a Bayesian analysis of the data for the narrower reference class. If analysed within Bayesian statistics, samples of frequencies in narrower reference classes lead to a posterior distribution for $freq(T|R')$. Contrary to Kyburg and Teng's approach to direct inference, not every point in an interval is treated equally (see Sect. 2).

Let D be a sequence of observations whether certain individuals in the narrower reference class belong to the target class and $l(D|freq(T|R') = v_i)$ the likelihood of these observations given the relative frequency of the target class in R' is v_i. Then

---

[11] A sample is random if and only if in each draw of the sample every member of the population has the same probability of entering the sample.

$$PROB(freq(T|R') = v_i|D) = \frac{1}{T} \times p_i^* \times l(D|freq(T|R') = v_i), \tag{3}$$

where $T = \sum_{i=1}^{n} p_i^* \times l(D|freq(T|R') = v_i)$ is a normalizing constant.

The direct inference probability is the expected value of the posterior distribution:

$$PROB(c \in T) = E[freq(T|R')] = \sum_{i=1}^{n} v_i \times PROB(freq(T|R') = v_i|D). \tag{4}$$

Modulo prior distribution, the posterior distribution accounts for the fact that different values for $freq(T|R')$ explain that we observe a particular sample to a different degree. In our example, $freq(T|R') = 0.5$ explains the fact that we observed 8 T-individuals much better than $freq(T|R') = 0.1$. Hence, modulo prior probability, the posterior probability of $freq(T|R') = 0.5$ is much higher than the posterior probability of $freq(T|R') = 0.1$. Bayesian statistics has always been subject to the criticism that the posterior probability is subjective, because it strongly depends on the prior distribution chosen. However, since it contains the information of the frequency of the target class in the broader reference class, the natural distribution of sub-reference classes in the broader reference class is a reasonable prior distribution for $freq(T|R')$.

Equation (3) captures core intuitions about the conflict of narrowness and precision. The frequency in the broader reference class will influence the direct inference probability, if there is only a small sample for the narrower reference class available (this is the case in which the estimate for the narrower reference class is rather imprecise). As the sample size increases, the frequency in the broader reference class will loose influence on the direct inference probability. Again, a Bayesian line of reasoning is not viable in Thorn's and Pollock's approach. The low variance of the distribution for subsets in broader reference classes will make it almost impossible to update the prior on basis of data directly about the narrower reference class.

5 Conclusions and Future Work

Kyburg and Teng's approach leads to unreasonable direct inference probabilities. Thorn's approach and Pollock's approach are more promising. However, as my examples in Sect. 3.3 show, Thorn's Method 1 leads to unreasonable direct inference probabilities. For instance, if reference classes have sufficiently many members, then it leads to Kyburg and Teng's unreasonable Criterion of Precision.

The main reason for this is that Thorn considers arbitrary subsets of the broader reference class. However, for almost all subsets of the broader reference class it holds that the frequency of the target class is very close to the frequency of the target class in the broader reference class. Consequently, the probability distribution employed to build expected frequencies has too low variance. This point is more general and applies to any approach to direct inference that is based on combinatorial probabilities. In particular, it applies to Pollock's approach to direct inference.

In addition, Thorn's Method 1 is of limited practical applicability to diagnosis and prediction in the empirical sciences. It is silent about the case in which a sample of the target class in the narrower reference class is available. These samples lead to statistical estimates for relative frequencies of the target class in the narrower reference class.

In response to these two shortcomings, I developed a new Bayesian solution to the conflict of narrowness and precision that is based on two main assumptions. First, to

determine expected values, instead of the distribution of frequencies in arbitrary sub-classes, the *natural distribution* should be employed, i.e., the distribution of frequencies in sub-reference classes of the broader reference class that we actually use in direct inference should be employed. Second, probabilities obtained by the natural distribution need to be aggregated with estimates for the frequencies of the target class in the narrower reference class obtained from data. The resulting approach equates the direct inference probability with the expected value of the posterior distribution in the narrower reference class. A reasonable prior is given by the natural distribution. However, further research is needed to determine the relevant natural distribution, i.e., to determine the p_i^*.

Appendix

Let V and $W := \{w_1, \ldots, w_n\} \subseteq \mathbb{R}$ be finite sets. Let P be a probability function on V. Then a function $X : V \mapsto W$ is called a discrete random variable. The distribution of X is given by $P(X = w) := P(\{v \in V | X(v) = w\})$ for all $w \in W$. Its expected value and its variance are given by

$$E[X] := \sum_{i=1}^{n} W_i \cdot P(X = W_i), Var[X] := \sum_{i=1}^{n} (W_i - E[X])^2 \cdot P(X = W_i).$$

The following facts are well known in combinatorics (for instance, Larsen and Marx 2012, p.110, p.143, pp.191–192).

Theorem 1 *Let $|R| = N$, $|R'| = k$ and $freq(T|R) = r$.*
Then $X_{Freq}: POW(R, |R'|) \mapsto \{\max\{0, ((|R'| + rN) - N)\}, \ldots, \min\{|R'|, rN\}\}: X_{Freq}(S) = |T \wedge S|$ is hypergeometrically distributed with parameters N, rN, k. I.e.,

$$PROB(X_{Freq} = s) = \frac{\binom{rN}{sk}\binom{N - rN}{k - sk}}{\binom{N}{k}}.$$

The expected value of X_{Freq} is kr and therefore $E[freq(T|S)] = E[\frac{|T \wedge S|}{|S|}] = \frac{1}{k}E[|T \wedge S|] = r$.
The variance of X_{Freq} is $kr(1 - r)\frac{N-k}{N-1}$. Hence, $Var[freq(T|S)] = \frac{1}{k^2}Var[|T \wedge S|] = \frac{r(1-r)\frac{N-k}{N-1}}{k}$.

Hence,

$$z_i = \frac{\binom{rN}{v_i k}\binom{N - rN}{k - v_i k}}{\binom{N}{n}}, p_i = \frac{\binom{rN}{v_i k}\binom{N - rN}{k - v_i k}}{\sum_{j=1}^{n}\binom{rN}{v_j k}\binom{N - rN}{k - v_j k}}. \tag{5}$$

Here is a Matlab-code to calculate Thorn's expected frequencies:

```matlab
r=0.25 %freq(T|R)
k=100% |R'|
N=1000 %|R|
V=[0.2:0.01:0.4] %vector of the set of possible values for freq(T|R')=v_i
M0=V*k; %vector of all possible values for freq(T&R')
M1=int16(M0);
z=zeros(1,length(V));
for i=1:length(M1)
    z(i)=hygepdf(M1(i),N,r*N,k);
end
z %Vector of the z_i
p=z/sum(z) %Vector of the p_i
s=p;
 for i=1:length(M1)
    s(i)=s(i)*V(i);
 end
 s %Vector of the p_i*z_i
Expfreq=sum(s) % Expected frequency:  E[freq(T|R')]
```

References

Kyburg, H. E., Jr. (1961). *Probability and the logic of rational belief*. Middletown: Wesleyan UP.

Kyburg, H. E., Jr., & Teng, C. M. (2001). *The theory of probability*. Cambridge: Cambridge University Press.

Larsen, R., & Marx, M. (2012). *An introduction to mathematical statistics and its applications*. London: Pearson.

Manolio, T., et al. (2009). Finding the missing heritability of complex disease. *Nature, 461*(7265), 1564–1570.

Paris, J. B., Watton, P. N., & Wilmers, G. M. (2000). On the structure of probability functions in the natural world. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 8*(03), 311–329.

Pollock, J. L. (1990). *Nomic probability and the foundations of induction*. Oxford: Oxford University Press.

Pollock, J. L. (2011). Reasoning defeasibly about probabilities. *Synthese, 181*(2), 317–352.

Reichenbach, H. (1949). *The Theory of probability*. Berkeley: University of California Press.

Salari, K., Hugh, W., & Ashley, E. (2012). Personalised medicine: Hope or hype? *European Heart Journal, 33*, 1564–1570.

Stone, M. (1987). Kyburg, Levi, and Petersen. *Philosophy of Science, 54*(2), 244–255.

Thorn, P. D. (2017). On the preference of more specific reference classes. *Synthese, 194*(6), 2025–2051.

Thorn, Paul D. (2012). Two problems of direct inference. *Erkenntnis, 76*(3), 299–318.

Venn, J. (1888). *The Logic of Chance* (3rd ed.). London: Macmillan.

Wallmann, C., & Williamson, J. (2017). Four approaches to the reference class problem. In L. Wronski & G. Hofer-Szabo (Eds.), *Making it formally explicit: Probability, causality, and determinism* (pp. 61–81). Berlin: Springer.

Williamson, J. (2013). Why frequentists and Bayesians need each other. *Erkenntnis, 78*(2), 293–318.