The Variety-of-Evidence Thesis and the Reliability of Instruments: A Bayesian-Network Approach

Stephan Hartmann[1] and **Luc Bovens**[2]
[1] University of Pittsburgh, Center for Philosophy of Science, Pittsburgh, PA 15260, USA
email: shart@pitt.edu
[2] University of Colorado at Boulder, Department of Philosophy, Boulder CO 80309, USA
email: bovens@spot.colorado.edu

Abstract

The variety of evidence thesis in confirmation theory states that more varied supporting evidence confirms a hypothesis to a greater degree than less varied evidence. Under a very plausible interpretation of this thesis, positive test results from multiple independent instruments confirm a hypothesis to a greater degree than positive test results from a single instrument. We invoke Bayesian Networks to model confirmation on grounds of evidence that is obtained from less than fully reliable instruments and show that the variety of evidence thesis is not sacrosanct when testing is conducted with less than fully reliable instruments: under certain conditions, a hypothesis receives more confirmation from evidence that is obtained from one rather than from more independent instruments. In the appendix, we prove certain convergence results for large numbers of positive test results from single versus multiple less than fully reliable instruments.

1 Introduction

In the Bayesian tradition in philosophy of science, problems are often framed in a highly idealized format. The price of relaxing these idealizations is mathematical complexity. In the late 70s, an axiomatic approach to conditional independence was developed within a Bayesian framework. This approach in conjunction with developments in graph theory are the two pillars of the theory of Bayesian Networks, which is a theory of probabilistic reasoning in artificial intelligence (e.g. [Pearl, 1988]). The theory has been very successful over the last two decades and has found a wide array of practical applications. Aside from work in the theory of causation, philosophers have been sadly absent in reaping the fruits from these new developments in artificial intelligence. This is unfortunate, since there are some questions in philosophy of science in which the route to progress has been blocked by a type of complexity that is precisely the type of complexity that Bayesian Networks are designed to deal with questions in which there are multiple variables in play and the conditional independences between these variables can be clearly identified. We will assess one such question in confirmation theory. It is often said that it is better to have more rather than less varied evidence for the confirmation of a hypothesis and philosophers have attempted to give a Bayesian account of this phenomenon. It requires some elucidation what it means for evidence to be more varied. One obvious interpretation is that evidence which comes from multiple independent test instruments is more varied than evidence that comes from a single test instrument. Suppose that our test results come from less than fully reliable (LTFR) instrument, as is often the case in scientific experiments. (cf. [Franklin, 1986], pp. 165-191.) We will show that the variety of evidence thesis is not sacrosanct under this interpretation. Depending on a range of relevant parameters, more varied evidence, *in casu* evidence that comes from multiple instruments rather than from a single instrument, may or may not confirm the hypothesis to a greater degree.

2 Confirmation with LTFR Instruments

Consider a very simple scenario. Let there be a hypothesis, a (test) consequence of the hypothesis, a LTFR instrument and a report from the LTFR instrument to the effect that the consequence holds or not. To model this scenario, we need four propositional variables (written in italic script) and their values (written in roman script):

(1) HYP can take on two values: HYP, i.e. the hypothesis is true and $\overline{\mathrm{HYP}}$, i.e. the hypothesis is false;

(2) CON can take on two values: CON, i.e. the consequence holds and $\overline{\mathrm{CON}}$, i.e. the consequence does not hold;

(3) REL can take on two values: REL, i.e. the instrument is reliable and $\overline{\mathrm{REL}}$, i.e. the instrument is not reliable;

(4) REP can take on two values: REP, i.e. there is a confirming report, or, in other words, a report to the effect that the consequence holds, and $\overline{\mathrm{REP}}$, i.e. there is a disconfirming report, or, in other words, a report to the effect that the consequence does not hold.

A probability distribution over these variables contains 2^4 entries. The number of entries will grow exponentially with the number of propositional variables. To represent the information in a more parsimonious format, we construct a Bayesian Network.

A Bayesian Network organizes the variables into a *Directed Acyclical Graph* (DAG), which encodes a range of

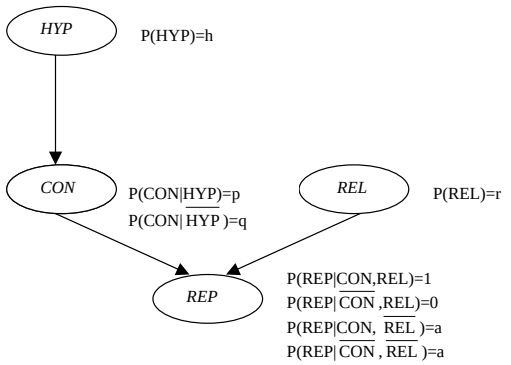

Figure 1: The basic model

(conditional) independences. There is a certain heuristic that governs the construction of the graph: There is an arrow between two nodes iff the variable in the parent node has a *direct influence* on the variable in the child node. In the case at hand, whether the consequence holds is directly influenced by and only by whether the hypothesis is true or not; whether there is a report to the effect that the consequence holds is directly influenced by and only by whether the consequence holds or not and whether the instrument is reliable or not. Hence, we construct the basic graph in figure 1 in which the node with the variable HYP is a parent node to the node with the variable CON and the nodes with the variables CON and REL are in turn parent nodes to the node with the variable REP.

We stipulate probability distributions for the variables in the root nodes of the graph.

$$P(\mathrm{HYP}) = h, \quad P(\mathrm{REL}) = r \tag{1}$$

with $0 < h, r < 1$, and conditional probability distributions for the variables in the other nodes given any combination of values of the variables in their respective parent nodes. Consider the node with the variable CON which is a child node to the node with the variable HYP. We take a broad view of what constitutes a consequence, that is, we do not require that the truth of the hypothesis is either a necessary or a sufficient condition for the truth of the consequence. Rather, a consequence is to be understood as follows: The probability of the consequence given that the hypothesis is true is greater than the probability of the consequence given that the hypothesis is false:

$$P(\mathrm{CON}|\mathrm{HYP}) = p > q = P(\mathrm{CON}|\overline{\mathrm{HYP}}) \tag{2}$$

Consider the node with the variable REP, which is a child node to the nodes with the variables CON and REL. How can we model the workings of an unreliable instrument? Let us make an idealization: We suppose that we do not know whether the instrument is reliable or not, but if it is reliable, then it is fully reliable and if it is not reliable, then it is fully unreliable. Let a fully reliable instrument be an instrument that provides maximal information: It is an instrument that says of what is that it is, and of what is not that it is not:

$$P(\mathrm{REP}|\mathrm{REL}, \mathrm{CON}) = 1, \; P(\mathrm{REP}|\mathrm{REL}, \overline{\mathrm{CON}}) = 0 \tag{3}$$

Let a fully unreliable instrument be an instrument that provides minimal information: It is an instrument that is no better than a randomizer:

$$P(\mathrm{REP}|\overline{\mathrm{REL}}, \mathrm{CON}) = P(\mathrm{REP}|\overline{\mathrm{REL}}, \overline{\mathrm{CON}}) = a \tag{4}$$

with $0 < a < 1$. Let us call a the randomization parameter. We can now construct the Bayesian Network by adding the probability values to the graph in figure 1.

What's so great about Bayesian Networks? A central theorem in the theory of Bayesian Networks states that a joint probability distribution over any combination of values of the variables in the Network is equal to the product of the probabilities and conditional probabilities for these values as expressed in the Network. For example, suppose we are interested in the joint probability of $\mathrm{HYP}, \overline{\mathrm{CON}}, \mathrm{REP}$ and $\overline{\mathrm{REL}}$. We can read the joint probability directly off of figure 1: $P(\mathrm{HYP}, \overline{\mathrm{CON}}, \mathrm{REP}, \overline{\mathrm{REL}}) = P(\mathrm{HYP})P(\overline{\mathrm{CON}})P(\overline{\mathrm{CON}}|\mathrm{HYP})P(\mathrm{REP}|\overline{\mathrm{REL}}, \overline{\mathrm{CON}}) = h(1-r)(1-p)a$. Standard probability calculus teaches us how to construct marginal distributions out of joint distributions and subsequently conditional distributions out of marginal distributions. When implemented on a computer, Bayesian Networks provide a direct answer to such queries.

We are interested in the probability of the hypothesis given that there is a report from a LTFR instrument that the consequence holds. This probability is $P^*(\mathrm{HYP}) = P(\mathrm{HYP}|\mathrm{REP}) = P(\mathrm{HYP}, \mathrm{REP})/P(\mathrm{REP})$. For ease of representation, we will abbreviate $1 - x$ as $\overline{x}$.

$$P^*(\mathrm{HYP})) = \frac{h(pr + a\overline{r})}{hr(p-q) + qr + a\overline{r}} \tag{5}$$

We measure the degree of confirmation that the hypothesis receives from a confirming report by the difference:

$$P^*(\mathrm{HYP}) - P(HYP) = \frac{h\overline{h}(p-q)\, r}{hr(p-q) + qr + a\overline{r}} \tag{6}$$

We know now how to model the degree of confirmation that a hypothesis receives from a single confirming report concerning a single consequence of the hypothesis by means of a single LTFR instrument. This basic model will be the paradigm to model complex strategies to improve the degree of confirmation that can be obtained from LTFR instruments.

3 Repeated Testing

Suppose that we have tested a single consequence of the hypothesis by means of a single LTFR instrument. We have received a confirming report, but we want to have additional confirmation for our hypothesis. We might want to run more tests of the very same consequence. Now there are two possibilities. Either we can take our old LTFR instrument and run the test a couple more times. Or we can choose new and independent LTFR instruments and test the very same consequence with these new instruments. We are curious to know which strategy for the confirmation of the hypothesis is the better strategy assuming that we do indeed receive more reports to the effect that the consequence hold. In other words, which strategy yields a higher degree of confirmation?

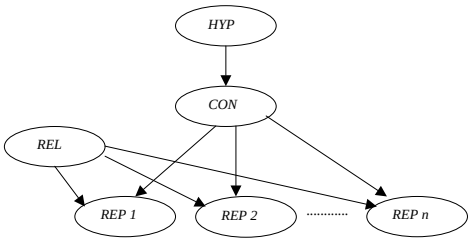

Figure 2: Repeated testing of the same consequence with

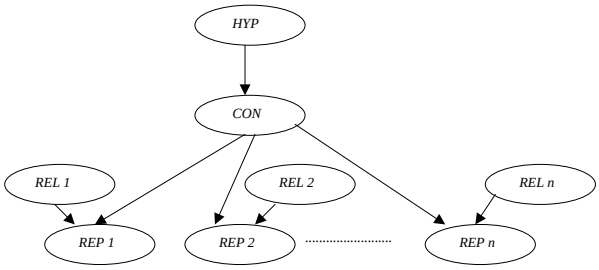

Figure 3: Repeating testing of the same consequence with multiple instruments

Is there an univocal answer to this question, or is one strategy more successful under certain conditions, while the other strategy is more successful under other conditions? [Bovens and Hartmann, 2001] have investigated the case for one additional test report, either from the same or from different LTFR instruments. Here we extend this research by investigating the general case of n test reports and prove certain convergence results.

Let us first model the degree of confirmation that the hypothesis receives from additional confirming reports from the same LTFR instrument. In figure 2, we add additional nodes to our basic graph to represent the binary variables $REP_2, \ldots, REP_n$, and substitute REP_1 for REP. Just like REP_1, it is also the case that $REP_2, \ldots, REP_n$ are directly influenced by REL and CON and so $2(n-1)$ more arrows are drawn in. We impose a condition of symmetry on the probability distribution P for this graph: Also for these additional reports the instrument is either fully reliable or it is fully unreliable with the same randomization parameter a.

Second, we model the degree of confirmation that the hypothesis receives from an additional confirming reports from $n-1$ other independent LTFR instrument. In figure 3, we add additional nodes to our basic graph for the variables $REL_2, \ldots, REL_n$, which express whether the other instruments are reliable or not, and add additional nodes for the variables $REP_2, \ldots, REP_n$ which express whether the other instruments provide reports to the effect that the consequence holds or not. REP_i is directly influenced by REL_i and CON for $i = 2, \ldots, n$: We draw in two more arrows for each pair of variables $\{REP_i, REL_i\}$. To keep matters simple, we impose a condition of symmetry on the probability distribution P' for this graph: There is an equal chance r that the instruments are reliable and if the instruments are unreliable then they randomize at the same level a. To compare the scenario with one instrument to the scenario with n instruments we need to impose a ceteris paribus condition: For this reason we postulate the same values h, p, q, r and a for the probability distributions P and P'.

The instruments are independent of one another. What this means is that

$$REP_i \perp REP_1, \ldots, REP_{i-1}, REP_{i+1}, \ldots, REP_n | CON \tag{7}$$

for all $i = 1, \ldots, n$ ($\perp$ represents the conditional independence relation.).

Suppose that we know that the consequence holds or we know that the consequence does not hold. Then there is a certain chance that we will receive a report to the effect that the consequence holds. Now whether we receive other reports to this effect or not, does not affect this chance. An independent instrument may not always provide us with an accurate report, but it is not influenced by what other instruments report. It can be shown by standard techniques in the theory of Bayesian Networks that (7) is a conditional independence that can be read off from the graph in figure 3.

We turn to the question whether, *ceteris paribus*, the hypothesis receives more confirmation from other confirming reports from one and the same LTFR instrument or from more independent LTFR instruments. We follow our standard procedure and calculate the difference $\Delta P = P'(\mathrm{HYP}|\mathrm{REP}_1, \mathrm{REP}_2, \ldots, \mathrm{REP}_n) - P(\mathrm{HYP}|\mathrm{REP}_1, \mathrm{REP}_2, \ldots, \mathrm{REP}_n)$. Aside from positive definite scaling factors, ΔP is given by

$$\Delta P \propto h\overline{h}(p-q)r\overline{r}\left(\sum_{k=0}^{n-1}\binom{n}{k} r^{n-k-1}(a\overline{r})^k - \overline{r}^{n-1}\right) \tag{8}$$

This expression allows us to construct phase curves which separate the parameter space in subspaces with $\Delta P > 0$ and $\Delta P < 0$.

The graph in figure 4 represents this inequality for n instruments with $n = 2, 8, 14$, and 20. For values of a and r above the phase curve, $\Delta P > 0$, i.e. confirming reports from more instruments provide more confirmation to the hypothesis; for values of a and r on the phase curve, $\Delta P = 0$, i.e. it does not make any difference whether we receive positive reports from one or more instruments; for values of a and r below the phase curve, $\Delta P < 0$, i.e. confirming reports from one instrument provide more confirmation to the hypothesis. Furthermore, the phase curve intersects the r-axis for $r_0 = .5$ and the a-axis for $a_0^{(n)} = \sqrt[n-1]{1/n}$ for any values of n. Note that $\lim_{n\to\infty} a_0^{(n)} = 1$.

Do these results seem plausible at some intuitive level? There are two conflicting intuitions at work here. On the one hand, we are tempted to say that confirming results from more instruments is the better way to go, since independence is a good thing. On the other hand, if we receive consistent confirming reports from a single instrument, then we feel more

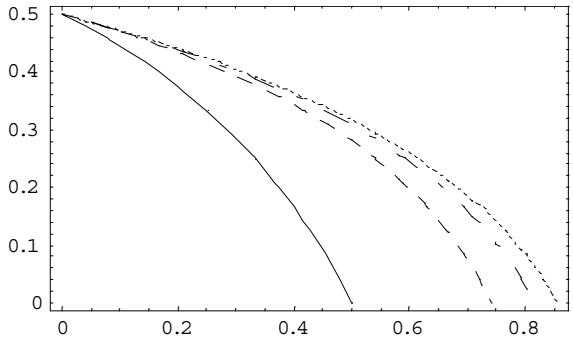

Figure 4: Phase curves for $n = 2$ (full line), $n = 8$ (dashed line), $n = 14$ (dot-dashed line), $n = 20$ (dotted line) for repeated testing of the same consequence for the parameters a and r: n instruments provide a higher degree of confirmation for the hypothesis above the curves, one instrument provides a higher degree of confirmation underneath the curves.

confident that the instrument is not a randomizer and this increase in confidence in the reliability of the instrument benefits the confirmation of the hypothesis. For higher values of r, the former consideration becomes more weighty than the latter: There is not much gain to be made anymore in our confidence in the reliability of the instrument(s) and we might as well enjoy the benefits of independence. For lower values of a, the latter consideration becomes more weighty: If we are working with an instrument which, if unreliable, has a low chance of providing confirming reports, then consistent confirming reports constitute a substantial gain in our confidence in its reliability, which in turn benefits the confirmation of the hypothesis. Furthermore, as there are more and more test reports in play, the latter consideration becomes more weighty, also for higher values of a: when we have little confidence in our instruments, one instrument yielding n confirming test results is more beneficial for the confirmation of the hypothesis than n instruments yielding n confirming results for low to high (but not extremely high) values of the randomization parameter.

4 Coherent Consequences

Another strategy to raise the degree of confirmation for a hypothesis is to identify a range of consequences which all can be assessed by a single or by multiple independent LTFR instruments. Again, we extend the results for two consequences in [Bovens and Hartmann, 2001] to n consequences and prove convergence results. Following our heuristic, the hypothesis (HYP) directly influences the consequences (CON_i) for $i = 1, 2, \ldots$. Figure 5 represents the scenario in which there is a single instrument: Each consequence (CON_i) conjoint with the reliability of the single instrument (REL) directly influences the report about the consequence in question (REP_i). Figure 6 represents the scenario in which there are more independent instruments: Each consequence (CONi) conjoint with the reliability of the instrument that tests this consequence (REL_i) directly influences the report

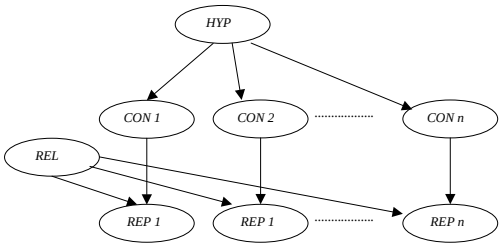

Figure 5: Testing of multiple consequences with a single instrument

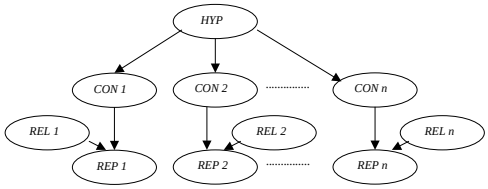

Figure 6: Testing of multiple consequences with multiple instruments

about the consequence in question (REP_i). We define a probability distribution P for the DAG in figure 5 and a probability distribution P' for the DAG in figure 6. We impose the *symmetry condition* within each distribution and the *ceteris paribus* condition between distributions for all the relevant parameters.

Does the hypothesis receive more confirmation from n confirming reports from one and the same LTFR instrument or from independent LTFR instruments, *ceteris paribus*. Let's follow our standard procedure and calculate $\Delta P = P'(HYP|REP_1, REP_2, \ldots, REP_n) - P(HYP|REP_1, REP_2, \ldots, REP_n)$. Aside from a positive definite scaling factor, ΔP is given by

$$\Delta P \propto (pr + a\overline{r})^n (q^n r + a^n \overline{r}) - (qr + a\overline{r})^n (p^n r + a^n \overline{r}) \quad (9)$$

Again, this expression allows us to construct the relevant phase curves. To evaluate equation (9), we first assume that the tests are reasonably strong by fixing $p = .9$ and $q = .1$ and construct a phase curves for values of a and r in figure 7 for $n = 2, 8, 14$, and 20. The general characteristics of these curves can be understood analyically. It can be shown analytically that the phase curve intersects the a-axis at

$$a^* = \sqrt[n-1]{\frac{1}{n}\,\frac{p^n - q^n}{p - q}} \quad (10)$$

and converges, for increasing r, at

$$a^{**} = \sqrt[n-1]{n\,\frac{p - q}{p^n - q^n}} \;<\; a^* \,. \quad (11)$$

Note that $\lim_{n\to\infty} a^* = p$ and $\lim_{n\to\infty} a^{**} = q$ (see appendix A).

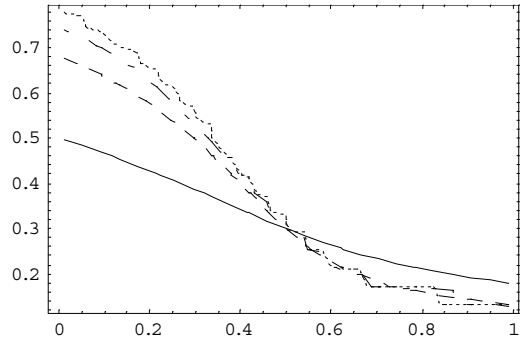

Figure 7: Phase curves for $n = 2$ (full line), $n = 8$ (dashed line), $n = 14$ (dot-dashed line), $n = 20$ (dotted line) for testing of multiple consequences for the parameters r and a with $p = .9$ and $q = .1$: n instruments provide a higher degree of confirmation for the hypothesis above the curves, one instrument provides a higher degree of confirmation underneath the curves.

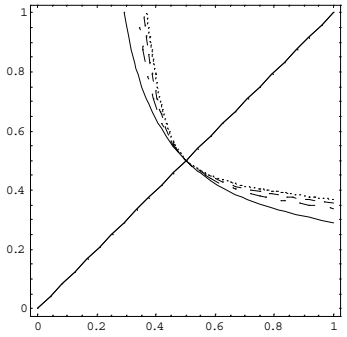

Figure 8: Phase curves for $n = 2$ (full line), $n = 8$ (dashed line), $n = 14$ (dot-dashed line), $n = 20$ (dotted line) for testing of multiple consequences for the parameters p and q with $a = .5$ and $r = .7$: One instrument provides a higher degree of confirmation for the hypothesis above the curves, multiple instruments provide a higher degree of confirmation underneath the curves.

Secondly, we set the values for the randomization and the reliability parameters at $a = .5$ and $r = .7$ and construct phase curves for values of p and q for $n = 2, 8, 14$, and 20 in figure 8. Examining the phase curves for other values of a and r, we can make the following observations: (i) The intersection point of the phase curves with the diagonal is always at (a, a); (ii) If we set $r = .5$, then the curve is the same for any value of n, as we have shown in the appendix; (iii) For $r > .5$, the curves for higher n are above the curves for lower n, while for $r < .5$, the curves for higher n are below the curves for lower n.

Whether it is better for the hypothesis to receive confirming results about multiple consequences from a single or from multiple instruments is contingent on the precise values of the parameters a, r, p, and q. The interpretation of figure 7 is similar to our interpretation of figure 4. But we can make some interesting additional observations. Note that if the value of p exceeds the randomization parameter a, then it is always better to receive confirming results from multiple instruments. On the other hand, if the randomization parameter a exceeds the value of q, then it is always better to receive confirming results from a single instrument. As to the interpretation of figure 8, we are only interested in the area below the straight line where $p > q$. We notice that if the q-value is set high, i.e. for weaker tests, one instrument tends to do better than more instruments. Why is this the case? The higher the q-values, the more likely the testable consequences will hold true and so coherent confirming reports will boost our confidence in the reliability of a single instrument even more. Hence higher q-values tend to favor a single instrument over multiple instruments. Furthermore, complex interaction effects between the number of reports n, the reliability parameter a and the randomization parameter r affect the threshold of the value of q at which confirming reports from a single instrument yields a higher degree of confirmation than multiple instruments.

5 Discussion: The Variety-of-Evidence Thesis

It is one of the textbook Bayesian success stories that an account can be provided of why variety of evidence is a good thing: The increment of confirmation that the hypothesis receives from confirming test results becomes smaller and smaller as we run the same old test over and again (e.g. [Earman, 1992], pp. 77-79 and [Howson and Urbach, 1993], pp. 119-123). The moral seems to be that it is better to get confirming results about one or more test consequences from a range of independent instruments than from one and the same instrument. We have shown that the picture is not quite so simple when we experiment with LTFR instruments. In some cases it is better to get confirming results about one or more test consequences from the same instrument than from multiple independent instruments.

(i) Figure 4 shows that if we are testing a single consequence, it is sometimes more beneficial to get confirming reports from the same instrument than from new instruments, *ceteris paribus*;

(ii) Figures 7 and 8 show that if we are testing different consequences, it is sometimes more beneficial to get confirming reports from the same instrument than from new instruments, ceteris paribus (section refs: 4).

By repeating the test with the same instrument, we gain confidence in the reliability of the instrument which benefits the confirmation of the hypothesis more than the independence of multiple instruments. Furthermore, when [Bovens and Hartmann, 2001] first established these results for two test reports, there was a lingering suspicion that the advantages of one LTFR instrument might wane for more than two test reports and that more benefits would accrue from using multiple independent instruments. This suspicion is false. Our convergence results show that the differential impact of testing with one or more LTFR instruments is no less pronounced as the number of test reports increases. We conclude that the variety of evidence thesis is not sacrosanct: Confirm-

ing reports from single rather than from multiple instruments about a single or multiple consequences may, *ceteris paribus*, provide more confirmation to a hypothesis when testing is carried out with LTFR instruments.

A Limits of a^* and a^{**}

Recall that a^* and a^{**} are given by

$$\begin{aligned} a^* &= \sqrt[n-1]{\frac{1}{n}\,\frac{p^n - q^n}{p - q}} \\ a^{**} &= \sqrt[n-1]{n\,\frac{p - q}{p^n - q^n}}\,. \end{aligned}$$

Let $q = xp$ with $0 < x < 1$. Then

$$\begin{aligned} a^* &= p\,\sqrt[n-1]{\frac{1}{n}\,\frac{1 - x^n}{1 - x}} \\ a^{**} &= q\,\sqrt[n-1]{n\,\frac{1 - x}{1 - x^n}}\,. \end{aligned}$$

Since

$$\sqrt[n-1]{\frac{1}{n}\,\frac{1 - x^n}{1 - x}} = \sqrt[n-1]{1 + x + \ldots + x^{n-1}} =: g(x)$$

and

$$\sqrt[n-1]{1 + 0 + \ldots + 0} \;\leq\; g(x) \;\leq\; \sqrt[n-1]{1 + 1 + \ldots + 1}$$

for $0 < x < 1$, it follows that $\lim_{n\to\infty} a^* = p$ and $\lim_{n\to\infty} a^{**} = q$.

References

[Bovens and Hartmann, 2001] Luc Bovens and Stephan Hartmann. Bayesian networks and the problem of unreliable instruments. Forthcoming in *Philosophy of Science*, 2001.

[Earman, 1992] John Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press, Cambridge, Mass., 1992.

[Franklin, 1986] Allan Franklin. *The Neglect of Experiment*. Cambridge University Press, Cambridge, 1986.

[Howson and Urbach, 1993] Colin Howson and Peter Urbach. *Scientific Reasoning - The Bayesian Approach*. Open Court, Chicago, 1993.

[Pearl, 1988] Judah Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, Calif., 1988.