Judgments of conditional probability *vs.* evidential support

An experimental comparison in accuracy and time-consistency

Rutgers, September 20, 2016

Vincenzo Crupi

Center for Logic, Language, and Cognition – University of Turin

Munich Center for Mathematical Philosophy – LMU

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–21

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12259

Judging the Probability of Hypotheses Versus the Impact of Evidence: Which Form of Inductive Inference Is More Accurate and Time-Consistent?

Katya Tentori,[a] Nick Chater,[b] Vincenzo Crupi[c,d]

[a]*Center for Mind/Brain Sciences, University of Trento*
[b]*Behavioural Science Group, Warwick Business School*
[c]*Center for Logic, Language, and Cognition, University of Turin*
[d]*Munich Center for Mathematical Philosophy, Ludwig Maximilian University*

Received 3 June 2014; received in revised form 20 December 2014; accepted 6 January 2015

Abstract

Inductive reasoning requires exploiting links between evidence and hypotheses. This can be done focusing either on the posterior probability of the hypothesis when updated on the new evidence or on the impact of the new evidence on the credibility of the hypothesis. But are these two cognitive representations equally reliable? This study investigates this question by comparing probability and impact judgments on the same experimental materials. The results indicate that impact judgments are more consistent in time and more accurate than probability judgments. Impact judgments also predict the direction of errors in probability judgments. These findings suggest that human inductive reasoning relies more on estimating evidential impact than on posterior probability.

outline

- background and motivation
- the experiment: methods and procedure
- results
- conclusions

Dan Osherson (Princeton)

The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity

Branden Fitelson[†‡]
University of Wisconsin–Madison

Contemporary Bayesian confirmation theorists measure degree of (incremental) confirmation using a variety of non-equivalent relevance measures. As a result, a great many of the arguments surrounding quantitative Bayesian confirmation theory are implicitly *sensitive to choice of measure of confirmation*. Such arguments are *enthymematic*, since they tacitly presuppose that certain relevance measures should be used (for various purposes) rather than other relevance measures that have been proposed and defended in the philosophical literature. I present a survey of this pervasive class of Bayesian confirmation-theoretic enthymemes, and a brief analysis of some recent attempts to resolve the problem of measure sensitivity.

1. Preliminaries.

1.1. Terminology, Notation, and Basic Assumptions. The present paper is concerned with the degree of incremental confirmation provided by evidential propositions E for hypotheses under test H, given background knowledge K, according to relevance measures of degree of confirmation c. We say that c is a *relevance measure* of degree of confirmation if and only if c satisfies the following constraints, in cases where E confirms, disconfirms, or is confirmationally irrelevant to H, given background knowledge K.[1]

[†]Department of Philosophy, University of Wisconsin, 600 North Park Street, Madison, WI 53706.

[‡]Thanks to Marty Barrett, Ellery Eells, Malcolm Forster, Ken Harris, Mike Kruse, Elliott Sober, and, especially, Patrick Maher for useful conversations on relevant issues.

1. I will not defend the *qualitative* Bayesian relevance notion of confirmation here (I will just *assume* it, as an underpinning for the *quantitative* issues I discuss below). Nor will I argue for the existence of a 'rational' probability function Pr of the kind required to give Bayesian confirmation theory its (objective) normative teeth. For a nice recent

Philosophy of Science, 66 (Proceedings) pp. S362–S378. 0031-8248/99/66supp-0028$0.00

Available online at www.sciencedirect.com
ScienceDirect
COGNITION
ELSEVIER
Cognition 103 (2007) 107–119
www.elsevier.com/locate/COGNIT

Comparison of confirmation measures ☆,☆☆

Katya Tentori [a,*], Vincenzo Crupi [a,c], Nicolao Bonini [a], Daniel Osherson [b]

[a] *CRD, DiSCoF, University of Trento, via Matteo del Ben 5, 38068 Rovereto (TN), Italy*
[b] *Department of Psychology, Princeton University, Green Hall, Washington Street, Princeton (NJ) 08540, United States*
[c] *Laboratory of Cognitive Psychology, CNRS & University of Aix-Marseille I, 3 Place Victor Hugo F-13331, Marseille, France*

On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues*

Vincenzo Crupi, Katya Tentori, and Michel Gonzalez†‡

Epistemologists and philosophers of science have often attempted to express formally the impact of a piece of evidence on the credibility of a hypothesis. In this paper we will focus on the Bayesian approach to evidential support. We will propose a new formal treatment of the notion of degree of confirmation and we will argue that it overcomes some limitations of the currently available approaches on two grounds: (i) a theoretical analysis of the confirmation relation seen as an extension of logical deduction and (ii) an empirical comparison of competing measures in an experimental inquiry concerning inductive reasoning in a probabilistic setting.

Philosophy of Science, 74 (April 2007) pp. 229–252. 0031-8248/2007/7402-0003$10.00

Average correlations for competing confirmation measures

measure (predictor)	average correlation with participants' confirmation judgments for the hypothesis: urn *A* selected	urn *B* selected
Z	.756*	.775*
L	.740*	.754*
N	.730*	.745*
M	.628†	.588
G	.549	.631
R	.619	.557
S	.594	.613
C	.586	.605
D	.573	.589
p(A[B] | e)	.488	.508

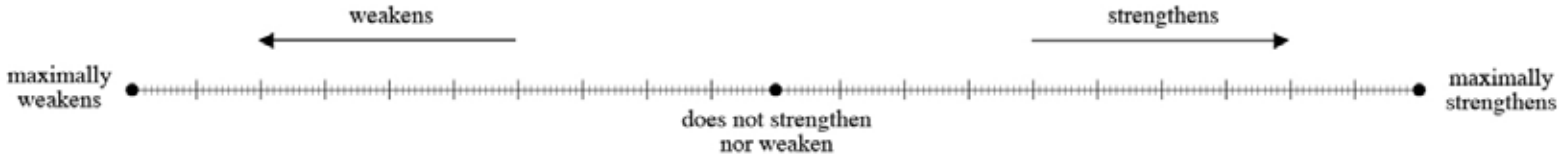
Reported values are the average of 26 correlations (one per participant) between the confirmation judgments predicted by each measure (on the basis of the randomness of the initial selection, the composition of the urns and the outcomes of previous draws) and the confirmation judgments expressed by each participant. Each correlation involved 10 observations.

$p(A[B] \mid e)$ denotes $p(A \mid e)$ or $p(B \mid e)$ as appropriate.

Comparisons by paired t-test with the average correlation for $p(A[B] \mid e)$:
*$p < .001$; †$p = .06$.

evidential impact (*confirmation*) scale

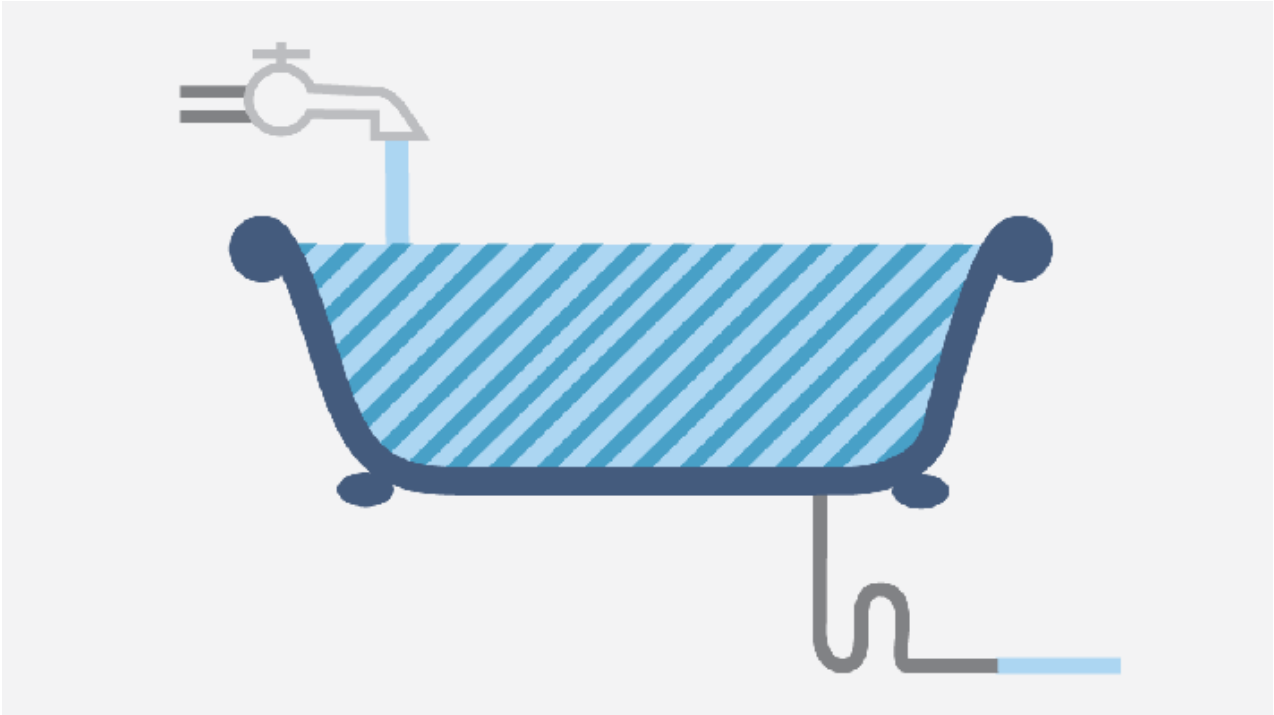
The information that [*e*]



the hypothesis that [*h*]

posterior probability *vs.* evidential support

stock and flow of credence…



posterior probability *vs.* evidential support

example:

a randomly selected student, *X*

h_1 = *X* owns a videogame console

h_2 = *X* likes going to the cinema

e = *X* is male

$Pr(h_1|e) = 0.69 \quad < \quad Pr(h_2|e) = 0.95$

$conf(h_1,e) > 0 \quad > \quad conf(h_2,e) = 0$

[participants also estimate $Pr(h_1) = 0.48$ on average]

posterior probability *vs.* evidential support

three measures:

$$L(h,e) = \frac{\Pr(e \mid h) - \Pr(e \mid \neg h)}{\Pr(e \mid h) + \Pr(e \mid \neg h)} \qquad \Rightarrow L(h_1,e) = 0.41$$

$$R(h,e) = \frac{\Pr(h \mid e) - \Pr(h)}{\Pr(h \mid e) + \Pr(h)} \qquad \Rightarrow R(h_1,e) = 0.18$$

$$Z(h,e) = \begin{cases} \dfrac{\Pr(h,e) - \Pr(h)}{\Pr(\neg h)} & \text{if } \Pr(h,e) \geq \Pr(h) \\ \dfrac{\Pr(h,e) - \Pr(h)}{\Pr(\neg h)} & \text{otherwise} \end{cases} \qquad \Rightarrow Z(h_1,e) = 0.40$$

the experimental questions

within subject consistency over time (test-retest reliability):

how reliable over time are judgments
of conditional probability vs. *evidential support?*

accuracy:

how accurate are judgments
of conditional probability vs. *evidential support?*

procedure and stimuli

preliminary phase

a convenience sample of 200 undergraduates (half male, half female) from the UCL population filled in a survey involving various personal questions such as:

– *do you have a driving license?*

– *can you dance?*

– *do you support any football team?*

– *do you own (at least) one videogame console?*

– *have you ever worked as a babysitter?*

…

procedure and stimuli

preliminary phase

a convenience sample of 200 undergraduates (half male, half female) from the UCL population filled in a survey involving various personal questions

responses were used to compute **frequency-based probabilities** – e.g., the probability of owning (at least) one videogame console in light of the evidence that one is male/female – and **corresponding values of evidential support** – the evidential impact of "*X* is male/female" on "*X* owns (at least) one videogame console" according to the likelihood ratio (*L*), the probability ratio (*R*), and the relative distance measure (*Z*)

procedure and stimuli

experimental stimuli

we constructed 56 pairs from

2 pieces of evidence (e = male, $\neg e$ = female) x 28 selected hypotheses

example

evidence: X is female

hypothesis: X owns (at least) one videogame console

procedure and stimuli

Appendix A: Arguments employed in the experiment

	$Pr(h\|e) > .5$ and $Pr(h\|\neg e) > .5$	$Pr(h\|e) > .5$ and $Pr(h\|\neg e) < .5$	$Pr(h\|e) < .5$ and $Pr(h\|\neg e) > .5$	$Pr(h\|e) < .5$ and $Pr(h\|\neg e) < .5$
$Imp(h,e) > 0$	e= X is a **male** h= X has a driving licence h= X owns (at least) one bike h= X can play volleyball e= X is a **female** h= X likes tea h= X likes carrots h= X likes shopping	e= X is a **male** h= X can play poker h= X supports a football team h= X likes beer h= X can play football h= X own (at least) a videogame console h = X can play basketball e= X is a **female** h= X likes ice-figure skating h= X likes candles h= X worked as a babysitter h= X own (at least) one cuddle toy h= X likes reading fashion magazines h= X can dance		e= X is a **male** h= X likes cigars h= X can surf h= X snores e= X is a **female** h= X owns (at least) one plant h= X has freckles h= X owns (at least) one weighting scale
$Imp(h,e) = 0$	e= X is a **male** h= X owns (at least) a mp3 player h= X likes going to the cinema e= X is a **female** h= X owns (at least) a mp3 player h= X likes going to the cinema			e = X is a **male** h = X has his/her own website h = X has (at least) 3 siblings e= X is a **female** h = X has his/her own website h = X has (at least) 3 siblings
$Imp(h,e) < 0$	e= X is a **female** h= X has a driving licence h= X owns (at least) one bike h= X can play volleyball e= X is a **male** h= X likes tea h= X likes carrots h= X likes shopping		e= X is a **female** h= X can play poker h= X supports a football team h= X likes beer h= X can play football h= X own (at least) a videogame console h= X can play basketball e= X is a **male** h= X likes ice-figure skating h= X likes candles h= X worked as a babysitter h= X own (at least) one cuddle toy h= X likes reading fashion magazines h= X can dance	e= X is a **female** h= X likes cigars h= X can surf h= X snores e= X is a **male** h= X owns(at least) one plant h= X has freckles h= X owns (at least) one weighting scale

Note: the crossed cells represent impossible combinations of probability and impact values.

procedure and stimuli

experimental stimuli

we constructed 56 pairs from
2 pieces of evidence (e = male, $\neg e$ = female) x 28 selected hypotheses

experimental phase

a new sample of 35 UCL undergraduates (mean age 22.43 years; 21 females) were recruited in the experiment

participants came twice to the laboratory, and on both occasions were asked to make 56 judgments of conditional probability and of evidential support

(to control for possible carry-over effects, the order of probability and confirmation questions was balanced across participants)

procedure and stimuli

evidential impact judgment (group 1, $n = 19$)

Consider a group of 200 students, 100 males and 100 females, from UCL.
Imagine we draw at random one of these 200 students. Let's call this student *A*.
Consider the following hypothesis (possibly true or false) concerning *A*:

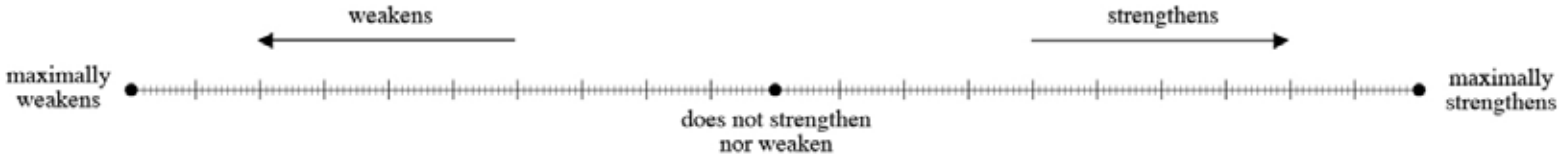
A owns (at least) one videogame console.

Now you are given a new piece of information (surely true) concerning *A*:

A is female.

How does this new piece of information (i.e., that *A* is male)
affect the hypothesis under consideration (i.e., that *A* owns at least one videogame console)?

The information that ***A is female***



the hypothesis that ***A owns (at least) one videogame console***

procedure and stimuli

probability judgment (group 1, $n = 19$)

Consider a group of 200 students, 100 males and 100 females, from UCL.

How many of the 100 *female* students *own (at least) one videogame console*? ____

procedure and stimuli

evidential impact judgment (group 2, $n = 16$)

Consider a group of 200 students, 100 males and 100 females, from UCL.
Imagine we draw at random one of these 200 students. Let's call this student *A*.
Consider the following hypothesis (possibly true or false) concerning *A*:

A owns (at least) one videogame console.

Now you are given a new piece of information (surely true) concerning *A*:

A is female.

How does this new piece of information (i.e., that *A* is female)
affect the hypothesis under consideration (i.e., that *A* owns at least one videogame console)?

Express your opinion by a number between –50 ("the information maximally
weakens the hypothesis") and +50 ("the information maximally strengthens the hypothesis").
Use 0 to indicate no impact at all ("the information does not weaken nor it strengthens
even a little the hypothesis").

The impact of the information that A ***is female*** on the hypothesis
that A ***owns (at least) one videogame console*** is: ____

procedure and stimuli

probability judgment (group 2, $n = 16$)

Consider a group of 200 students, 100 males and 100 females, from UCL.
Imagine we draw at random one of these 200 students. Let's call this student *A*.
Consider the following hypothesis (possibly true or false) concerning *A*:

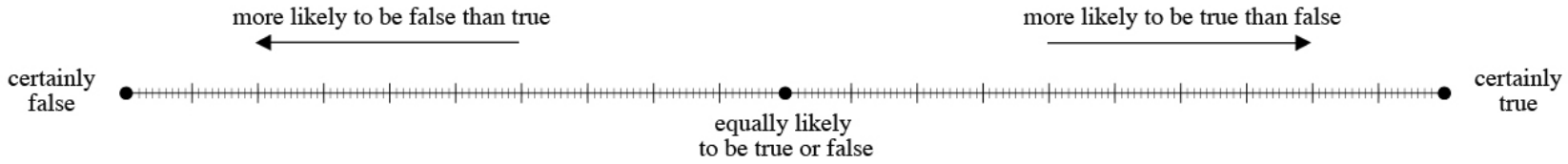
A owns (at least) one videogame console.

Now you are given a new piece of information (surely true) concerning *A*:

A is female.

In the light of this new piece of information (i.e., that *A* is female),
what is the probability of the hypothesis under consideration (i.e., that *A* owns
at least one videogame console)?

In the light of the information that *A* ***is female***,
the hypothesis that *A* **owns (at least) one videogame console** is



procedure and stimuli

experimental phase

a new sample of 35 UCL undergraduates (mean age 22.43 years; 21 females) were recruited in the experiment

participants came twice to the laboratory, and on both occasions were asked to make 56 judgments of conditional probability and of evidential support

(to control for possible carry-over effects, the order of probability and confirmation questions was balanced across participants)

participants were randomly divided in two groups, 19 were presented with a discrete probability scale and a continuous evidential impact scale, the other 16 were presented with a continuous probability scale and a discrete evidential impact scale

results

consistency over time

	average correlations	
participants	confirmation	probability
group 1	0.90	0.87
group 2	0.91	0.85
	n.s.	n.s.
all	0.91**	0.86

results

accuracy

	average correlations			
participants	*L*-conf	*Z*-conf	*R*-conf	probability
group 1	0.83**	0.80**	0.77**	0.67
group 2	0.82**	0.79**	0.76**	0.49
	n.s.	n.s.	n.s.	$p < 0.01$
all	0.82**	0.80**	0.77**	0.59

note: average correlations are higher for *L* than for *Z* and *R* ($p < 0.01$) and higher for *Z* than for *R* ($p < 0.01$)

results

accuracy

	average absolute error (on 0-100 scale)			
participants	*L*-conf	*Z*-conf	*R*-conf	probability
group 1	9,59	9.85	11.67	18.79
group 2	11.95	12.30	14.83	21.62
	$p < 0.05$	$p < 0.05$	$p < 0.05$	$p < 0.01$
all	10,67**	10.97**	13.12**	20.08

note: even if we selectively compare the *most* accurate probability judgments (frequency estimates in group 1) with the *least* accurate confirmation judgments (100-point scale in group 2) → $p < 0.01$

results

confirmation judgments predict probability errors

- for each participant, all the arguments (of the 56) with conf judged $\neq 0$
- compute the absolute difference between the corresponding survey *vs.* judged conditional probabilities, with (i) positive sign in case positive [negative] judged evidential impact + judged conditional probability overestimated [underestimated], and (ii) negative sign otherwise
- average all these differences (with sign) for each participant
- this index is positive if and only if the participant overestimated the conditional probability of the confirmed hypotheses and underestimated the conditional probability of the disconfirmed hypotheses *more than* s/he overestimated the conditional probability of the disconfirmed hypotheses and underestimated the conditional probability of the confirmed hypotheses
- the index is **strictly positive** for 80% of the participants
- mean for all is **+6.1** (on a 0-100 scale) (on average, systematic error of 6%)

summary

- evidential impact judgments are significantly more time-consistent and more accurate than corresponding conditional probability judgments
- evidential impact judgments can predict the direction of errors in probability judgments (when impact is positive [negative] people tended to overestimate [underestimate] the corresponding conditional probability)
- these results do not depend on the specific measure employed to quantify evidential impact, or on the specific scale used to collect the judgments

tentative conclusions, possible prospects

- a primitive kind of judgment?
- the reality-lab gap
- why?

thank you for your attention!