Fusion Confusion? Comments on Nancy Reid:
“BFF Four–Are we Converging?”

Deborah G. Mayo
The Fourth Bayesian, Fiducial and Frequentist Workshop (BFF4): Harvard University
May 2, 2017
I’m delighted to be part of a workshop linking statistics and philosophy of statistics! I thank the organizers for inviting me.

Nancy Reid’s “BFF Four–Are we Converging?” gives numerous avenues for discussion

She zeroes in on obstacles to fusion:

Confusion or disagreement on the nature of probability and its use in statistical inference
From Nancy Reid:

**Nature of probability**

*probability to describe physical haphazard variability*

- probabilities represent features of the “real” world in idealized form
- subject to empirical test and improvement
- conclusions of statistical analysis expressed in terms of interpretable parameters
- enhanced understanding of the data generating process

**probability to describe the uncertainty of knowledge**

- measures rational, supposedly impersonal, degree of belief given relevant information (Jeffreys)
- measures a particular person’s degree of belief, subject typically to some constraints of self-consistency ...
- often linked with personal decision-making
• As is common, she labels the second “epistemological”

But a key question for me is: what’s relevant for a normative epistemology, for an account of what’s warranted/unwarranted to infer
Reid quite rightly asks:

• in what sense are confidence distribution functions, significance functions, structural or fiducial probabilities to be interpreted?

• empirically? degree of belief?

• literature is not very clear
Reid:

We may avoid the need for a different version of probability by appeal to a notion of calibration (Cox 2006, Reid & Cox 2015)

• This is my central focus

I approach this indirectly, with analogy between philosophy of statistics and statistics
Carnap: Bayesians as Popper: Frequentists (N-P/Fisher)
Can’t solve induction but can build *logics of induction* or *confirmation theories* (e.g., Carnap 1962).

- Define a confirmation relation: $C(H, e)$ (, rather than $|$)
- logical probabilities deduced from first order languages
- to measure the ”degree of implication” or confirmation that $e$ affords $H$ (*syntactical*)
Problems

• Languages too restricted
• There was a continuum of inductive logics (tried to restrict via “inductive intuition”)
• How can a priori assignments of probability be relevant to reliability? (“guide to life”)
• Few philosophers of science are logical positivists, but the hankering for a logic of induction remains in some quarters
Popper:
“In opposition to [the] inductivist attitude, I assert that $C(H,e)$ must not be interpreted as the degree of corroboration of $H$ by $e$, unless $e$ reports the results of our sincere efforts to overthrow $H$.” (Popper 1959, 418)

“The requirement of sincerity cannot be formalized--“(ibid.)

“Observations or experiments can be accepted as supporting a theory (or a hypothesis, or a scientific assertion) only if these observations or experiments are severe tests of the theory–or in other words, only if they result from serious attempts to refute the theory.” (Popper 1994, 89)

-never successfully formulated the notion
Ian Hacking (1965) gives a logic of induction that does not require priors, based on (Barnard, Royall, Edwards)

“Law of Likelihood”: \( x \) support hypothesis \( H_1 \) more than \( H_0 \) if,

\[
\Pr(x;H_1) > \Pr(x;H_0) \text{ (i.e., if the likelihood ratio LR} > 1)\).

George Barnard, “there always is such a rival hypothesis viz., that things just had to turn out the way they actually did” (1972, 129).

\[
\Pr(\text{LR in favor of } H_1 \text{ over } H_0 ; H_0) = \text{ high.}
\]
Neyman-Pearson:

“In order to fix a limit between ‘small’ and ‘large’ values of [the likelihood ratio] we must know how often such values appear when we deal with a true hypothesis.” (Pearson and Neyman 1967, 106)

Sampling distribution of LR

A crucial criticism in statistical foundations
In statistics:

“Sampling distributions, significance levels, power, all depend on something more [than the likelihood function]—something that is irrelevant in Bayesian inference—namely the sample space.” (Lindley 1971, 436)

Once the data are in hand: Inference should follow the Likelihood Principle (LP):

In philosophy (R. Rosenkrantz defending the LP):

“The LP implies...the irrelevance of predesignation, of whether a hypothesis was thought of beforehand or was introduced to explain known effects.” (Rosenkrantz 1977, 122)

(don’t mix discovery with justification)
Probabilism vs Performance

• Are you looking for a way to assign degree of belief, confirmation, support in a hypothesis
  – considered epistemological

• Or to ensure long-run reliability of methods, coverage probabilities (via the sampling distribution)
  – considered only for long-run behavior, acceptance sampling
We require a third role:

• **Probativism** (severe-testing). To assess and control erroneous interpretations of data, post-data

The problems with selective reporting (Fisher) non-novel-data (Popper), are not problems about long-runs—

It’s that *we cannot say about the case at hand* that it has done a good job of avoiding the sources of misinterpretation.
Ian Hacking: “there is no such thing as a logic of statistical inference” (1980, 145)

Though I’m responsible for much of the criticism....
“I now believe that Neyman, Peirce, and Braithwaite were on the right lines to follow in the analysis of inductive arguments”

• Probability enters to qualify a claim inferred, it reports the method’s capabilities to control and alert us to erroneous interpretations (error probabilities)

• Assigning probability to the conclusion rather than the method is “founded on a false analogy with deductive logic” (Hacking, 141). – he’s convinced by Peirce
The only two who are clear on the false analogy:

**Fisher** (1935, 54):

“In deductive reasoning all knowledge obtainable is already latent in the postulates. ..The conclusions are never more accurate than the data. **In inductive reasoning** .. [t]he conclusions normally grow more and more accurate as more data are included. It should never be true, though it is still often said, that the conclusions are no more accurate than the data on which they are based.”

**Peirce** (“The probability of Induction” 1878):

“In the case of analytic [deductive] inference we know the probability of our conclusion (if the premises are true), but in the case of synthetic [inductive] inferences we only know the degree of **trustworthiness** of our proceeding.”
Neyman and His Performance

You could say Neyman gets his performance idea trying to clarify Fisher’s fiducial intervals.

• Neyman thought his confidence intervals were the same as Fisher’s fiducial intervals.

• In a (1934) paper (to generalize fiducial limits), Neyman said a confidence coefficient refers to “the probability of our being right when applying a certain rule” for making statements set out in advance. (623)

• Fisher was highly complimentary: Neyman “had every reason to be proud of the line of argument he had developed for its perfect clarity.” (Fisher comment in Neyman 1934, 618)
Neyman thinks he’s clarifying Fisher’s (1936, 253) equivocal reference to the “aggregate of all such statements...”. [1]

which involves, apart from \( \mu \), directly calculable quantities only, namely \( \bar{x} \) and \( s \), both of which are sufficient statistics, we may infer, without any use of probabilities a priori, a frequency distribution for \( \mu \) which shall correspond with the aggregate of all such statements as that made above, to the effect that the probability that \( \mu \) is less than \( \bar{x} - 2.145 \frac{s}{\sqrt{n}} \) is exactly one in forty.

“This then is a definite probability statement about the unknown parameter...” (Fisher 1930, 533)
It’s interesting too, to hear Neyman’s response to Carnap’s criticism of “Neyman’s frequentism”

**Neyman:**

“I am concerned with the term ‘degree of confirmation’ introduced by Carnap. ... [if] the application of the locally best one-sided test ...failed to reject the [test] hypothesis...“(Neyman 1955, 40)

*The question is: does a failure to reject the hypothesis confirm it?*

A sample $X = (X_1, ..., X_n)$ each $X_i$ is Normal, $N(\mu, \sigma^2)$, (NIID), $\sigma$ assumed known;

\[ H_0: \mu \leq \mu_0 \text{ against } H_1: \mu > \mu_0. \]

Test fails to reject $H_0$, $d(x_0) \leq c_\alpha$. 
Carnap says yes...

Neyman:
“....the attitude described is dangerous.
...the chance of detecting the presence [of discrepancy $\delta$ from $H_0$], when only [this number of] observations are available, is extremely slim, even if [$\delta$ is present].” (Neyman 1955, 41)

“The situation would have been radically different if the power function...were greater than...0.95.” (ibid.)

Merely surviving the statistical test is too easy, occurs too frequently, even when $H_0$ is false.
A post-data analysis is even better*:

Mayo and Cox 2006 (“Frequentist principle of evidence”):

“**FEV: insignificant result:** A moderate $P$-value is evidence of the absence of a discrepancy $\delta$ from $H_0$, only if there is a high probability $(1-c)$ the test would have given a worse fit with $H_0$ (i.e., $d(X) > d(x_0)$) were a discrepancy $\delta$ to exist.” (83-4)

$$\text{If } \Pr(\text{Pr}(d(X) > d(x_0); \mu = \mu_0 + \delta) \text{ is high} \quad d(X) \leq d(x_0);$$

*infer: any discrepancy from $\mu_0 < \delta$ [Infer: $\mu < \text{CI}_u$]*

(* severity for “acceptance”: Mayo & Spanos 2006/2011)
How to justify detaching the inference?

_Rubbing off_: The procedure is rarely wrong, therefore, the probability it is wrong in this case is low.

_What’s rubbed off?_

_(could be a probabilism or a performance)_

Bayesian epistemologists:

_(Having no other relevant information): A rational degree of belief or epistemic probability rubs off_

Attaching the probability to the claim differs from a report of well-testedness of the claim
Severe Probing Reasoning

The reasoning of the severe testing theorist is counterfactual:

\[ H: \mu \leq \bar{x}_0 + 1.96\sigma_x \]

\( (i.e., \mu \leq \text{CI}_u) \)

\( H \) passes severely because were this inference false, and the true mean \( \mu > \text{CI}_u \) then, very probably, we would have observed a larger sample mean.

(I don’t saddle Cox with my take, nor Popper)
How Well Tested (Corroborated, Probed) ≠ How Probable

We can build a “logic” for severity (it won’t be probability)

- both C and ~C can be poorly tested
- low severity is not just a little bit of evidence, but *bad or no evidence*
- Formal error probabilities *may* serve to quantify probativeness or severity of tests (for a given inference), they do not automatically give this-*must be relevant*
What Nancy Reid’s paper got me thinking about is the calibration point:

Here’s the longer quote:

“We may avoid the need for a different version of probability by appeal to a notion of calibration, as measured by the behaviour of a procedure under hypothetical repetition. That is, we study assessing uncertainty, as with other measuring devices, by assessing the performance of proposed methods under hypothetical repetition. Within this scheme of repetition, probability is defined as a hypothetical frequency.” (Reid and Cox 2015, 295)
Notions of calibration also vary!

(1) If we calibrate p-values by a Bayes factor or other probabilism, p-values exaggerate evidence

(2) If we calibrate Bayes factors by performance or severity they exaggerate what’s warranted to infer

“depends on one’s philosophy of statistics” Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman (2016, 342).
Notions of calibration also vary!

(1) If we calibrate p-values by a Bayes factor or other probabilism, p-values exaggerate evidence

(2) If we calibrate Bayes factors by performance or severity, they exaggerate what’s warranted to infer

“depends on one’s philosophy of statistics,” Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman (2016, 342).

Reid:
- it is unacceptable if a procedure yielding high-probability regions in some non-frequency sense are poorly calibrated

I agree. I take this as calling for the second (2), frequentist, calibration
This takes me to my last point: an irony about today’s ‘replication crisis’

In some cases it’s thought Big Data foisted statistics on fields unfamiliar with its dangers, and Reid discusses some foibles

A lot of consciousness-raising is going on

More hand-wringing than ever regarding cherry-picking, selection effects (p-hacking, significance seeking)

R.A. Fisher: “it’s easy to lie with statistics by selective reporting” (1955, p. 75)—new names, same problem
Returns to a question from back when the possibility of a logic of induction was still viable: can’t data speak for themselves?

Preregistration calls are everywhere: “Authors must decide the rule for terminating data collection before data collection begins and report this rule in the article.” (Simmons, Nelson, and Simonsohn 2011, 1362)

At the same time...

“Use of the Bayes factor gives experimenters the freedom to employ optional stopping without penalty. (In fact, Bayes factors can be used in the complete absence of a sampling plan...)” (Bayarri, Benjamin, Berger, Sellke 2016, 100)
What I take away from Nancy Reid’s talk is: if we don’t know what we mean by an account “works” we can’t tell how to calibrate
In the severe testing view:
In order for a calibration to be relevant to normative epistemology, that is to what is warranted to infer, (what’s well and poorly tested)

1. It must be directly affected by selection effects (cherry picking, multiple testing, stopping rules)

2. enable testing assumptions

3. enable statistical falsification.

Points to the need for further philosophical-statistical interaction
<table>
<thead>
<tr>
<th>Philosophy of Inductive/Statistical Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inductive Logics</td>
</tr>
<tr>
<td>Carnap C(H,e), Hacking</td>
</tr>
</tbody>
</table>

**Parallels in Formal Statistics (goes much further)**

| Bayesian and Likelihoodist accounts        | Fisherian, Neyman-Pearson frequentist methods: |
| Probability: to assign degree of confirmation, support, belief (posterior or comparative) | Probability: (a) to ensure reliable **performance** |
| **Probabilisms**                           | (b) severity of tests **probativeness** |
| *Fiducial?*                                | *Fiducial?* |

<32>
A complementary doctrine of Neyman violating equally the principles of deductive logic is to accept a general symbolical statement such as

\[ Pr\{ (\bar{x} - ts) < \mu < (\bar{x} + ts) \} = \alpha , \]

as rigorously demonstrated, and yet, when numerical values are available for the statistics \( \bar{x} \) and \( s \), so that on substitution of these and use of the 5 per cent. value of \( t \), the statement would read

\[ Pr\{ 92.99 < \mu < 93.01 \} = 95 \text{ per cent.} , \]

to deny to this numerical statement any validity. This evidently is to deny the syllogistic process
REFERENCES


