Consistent Estimators and the Argument from Inductive Risk

Corey Dethier Samuel C. Fletcher

Minnesota Center for Philosophy of Science
University of Minnesota, Twin Cities

April 7, 2023

NSF Grant No. 2042366

The classical debate

"As is well known, the acceptance or rejection of such a hypothesis presupposes that a certain level of significance or level of confidence or critical region be selected." (Rudner 1953, 3)

"the activity proper to the scientist is the assignment of probabilities (with respect to currently available evidence) to the hypotheses which, on the usual view, he simply accepts or rejects." (Jeffrey 1956, 237)

Three observations

1. *Strictly speaking*, Rudner's argument only applies to a (strict) NP testing framework. Fisherian/hybrid approaches don't employ acceptance levels or regions.

Three observations

1. *Strictly speaking*, Rudner's argument only applies to a (strict) NP testing framework. Fisherian/hybrid approaches don't employ acceptance levels or regions.

2. Modern rejoinders to Jeffrey—e.g., Douglas (2000), Steele (2013)—have focused on other ways that values can enter into the testing process or the need to communicate results.

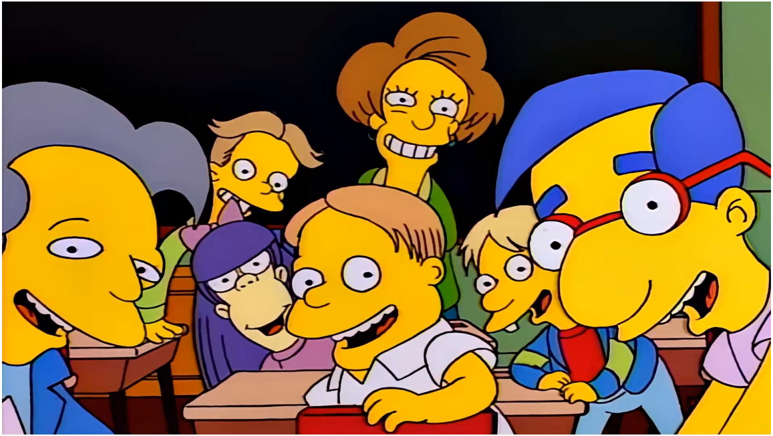
Three observations

1. *Strictly speaking*, Rudner's argument only applies to a (strict) NP testing framework. Fisherian/hybrid approaches don't employ acceptance levels or regions.

2. Modern rejoinders to Jeffrey—e.g., Douglas (2000), Steele (2013)—have focused on other ways that values can enter into the testing process or the need to communicate results.

3. It is thus an open question whether the scientist *qua (classical) statistician* must make value judgments.

Obligatory Simpsons reference



The thesis

Yes, the scientist qua statistician must make value judgments.

The thesis

Yes, the scientist qua statistician must make value judgments.

Focusing on the use of estimators:

- The choice of estimators is open to *exactly* analogous conditions as the choice of acceptance level.

The thesis

Yes, the scientist qua statistician must make value judgments.

Focusing on the use of estimators:

- The choice of estimators is open to *exactly* analogous conditions as the choice of acceptance level.
- All permissible estimators will eventually converge on the truth – a feature shared by Rudner's original example.

The thesis

Yes, the scientist qua statistician must make value judgments.

Focusing on the use of estimators:

- The choice of estimators is open to *exactly* analogous conditions as the choice of acceptance level.
- All permissible estimators will eventually converge on the truth – a feature shared by Rudner's original example.
- Which calls into question whether Rudner's example (a) can or (b) should play the role often assigned to it in the literature.

Estimators and inductive risk

Estimators

An **estimator** is a kind of test statistic: it's a rule for deriving a "best guess" (the "estimate") for a quantity of interest ("estimand") from the sample. E.g.:

1. If the quantity of interest is the population mean, the sample mean, median, or mode serve as the estimator.
2. Similarly for higher moments of the distribution, such as variance or skew.

Estimators

An **estimator** is a kind of test statistic: it's a rule for deriving a "best guess" (the "estimate") for a quantity of interest ("estimand") from the sample. E.g.:

1. If the quantity of interest is the population mean, the sample mean, median, or mode serve as the estimator.
2. Similarly for higher moments of the distribution, such as variance or skew.

Which estimator should we use?

Generally, prefer the estimator with the smallest (expected) loss.

Data	3.23	1.92	4.28	2.57	4.20	2.97	3.87	2.60
	2.62	3.72	2.72					

Loss function	**Estimator**	**Estimate**

Data are estimates for equilibrium climate sensitivity taken from Tokarska et al. (2020). Notably, there's discussion in the literature about how to estimate *variance* in this case (see Annan and Hargreaves 2011; Dethier 2022).

Which estimator should we use?

Generally, prefer the estimator with the smallest (expected) loss.

Data	3.23	1.92	4.28	2.57	4.20	2.97	3.87	2.60
	2.62	3.72	2.72					

Loss function	**Estimator**	**Estimate**
Mean absolute error $(\Sigma_x(\|Y - x\|)/n)$	Median	2.97

Data are estimates for equilibrium climate sensitivity taken from Tokarska et al. (2020). Notably, there's discussion in the literature about how to estimate *variance* in this case (see Annan and Hargreaves 2011; Dethier 2022).

Which estimator should we use?

Generally, prefer the estimator with the smallest (expected) loss.

Data	3.23	1.92	4.28	2.57	4.20	2.97	3.87	2.60
	2.62	3.72	2.72					

Loss function	**Estimator**	**Estimate**
Mean absolute error $(\Sigma_x(\lvert Y - x\rvert)/n)$	Median	2.97
Mean squared error $(\Sigma_x(Y - x)^2/n)$	Mean	3.15

Data are estimates for equilibrium climate sensitivity taken from Tokarska et al. (2020). Notably, there's discussion in the literature about how to estimate *variance* in this case (see Annan and Hargreaves 2011; Dethier 2022).

Which estimator should we use?

Generally, prefer the estimator with the smallest (expected) loss.

Data	3.23	1.92	4.28	2.57	4.20	2.97	3.87	2.60
	2.62	3.72	2.72					

Loss function	**Estimator**	**Estimate**
Mean absolute error ($\Sigma_x(\lvert Y-x\rvert)/n$)	Median	2.97
Mean squared error ($\Sigma_x(Y-x)^2/n$)	Mean	3.15
Mean quartic error ($\Sigma_x(Y-x)^4/n$)	unnamed	3.19

Data are estimates for equilibrium climate sensitivity taken from Tokarska et al. (2020). Notably, there's discussion in the literature about how to estimate *variance* in this case (see Annan and Hargreaves 2011; Dethier 2022).

Weighting different types of errors

These different functions represent different attitudes towards error.

MAE: errors of 1, 2, and 4 count for 1, 2, and 4.

MSE: errors of 1, 2, and 4 count for 1, 4, and 16.

MQE: errors of 1, 2, and 4 count for 1, 16, and 256.

Or: which estimator you should use depends on how you weight small vs. large errors.

Hypothesis testing

To carry out the most basic hypothesis test:

1. Calculate the value of the test statistic Z:

$$\frac{\text{estimator of the mean} - \text{hypothesized mean}}{\text{standard deviation}/\sqrt{\text{sample size}}}$$

2. Calculate the probability of observing a value greater than the result using a normal distribution.
3. The resulting probability is the *p-value*, which quantifies how well the evidence "fits" the hypothesis.

Hypothesis testing

To carry out the most basic hypothesis test:

1. Calculate the value of the test statistic Z:

$$\frac{\text{estimator of the mean} - \text{hypothesized mean}}{\text{standard deviation}/\sqrt{\text{sample size}}}$$

2. Calculate the probability of observing a value greater than the result using a normal distribution.
3. The resulting probability is the *p-value*, which quantifies how well the evidence "fits" the hypothesis.

Generalizing

The examples just given generalizes in a fairly trivial way to (almost all of?) the rest of inferential statistics, either by way of

1. relying on estimators—i.e., loss functions—to select an estimate (as here); or
2. relying explicitly or implicitly on loss functions in other ways (e.g., least-squares algorithms in regression).

So the same value-laden choices about loss functions are ubiquitous in classical statistics.

The parallel with Rudner's argument

Rudner's argument:

(P1) Scientists qua scientists must choose an acceptance level.

(P2) The choice of acceptance level requires weighting different errors.

(P3) Weighting different errors requires making value judgments.

∴ (C) Scientists qua scientists must make value judgments.

The parallel with Rudner's argument

Our argument:

(P1) Scientists qua classical statisticians must choose estimators.

(P2) The choice of estimator requires weighting different errors.

(P3) Weighting different errors requires making value judgments.

∴ (C) Scientists qua classical statisticians must make value judgments.

Consistent estimators and inductive risk

Consistency

In most contexts, any permissible estimator is *consistent*.

The main definition of consistency is that the estimator X_n "converges in probability" with the target θ:

$$\forall \epsilon > 0, \lim_{n \to \infty} Pr(|X_n - \theta| > \epsilon) = 0$$

Or, more simply:

$$\operatorname*{plim}_{n \to \infty} X_n(\theta) = \theta$$

Consistency and values

In the limit, the differences between the *estimates* generated by permissible estimators (almost surely) disappear.

Or: no matter how substantial the divergence of values, sufficient data will (almost surely) wash values out of the estimate.

And it has the same effect on Rudner's original example!

Consistency and acceptance levels

Let z_α indicate the "critical value": if the hypothesis is true, the probability of observing $|Z| > z_\alpha$ for a given sample size is α.

The accept and reject criteria are then:

Accept: if $|z| \leqslant z_\alpha$ **Reject:** $|z| > z_\alpha$

Claim: If the underlying estimator is consistent, as $n \to \infty$, the probability of accepting a true hypothesis goes to 1 and the probability of accepting a false one goes to 0 for any $z_\alpha \in (0, \infty)$.

Technically, you *could* define the acceptance region without reference to a test statistic. The resulting tests are not "completely consistent" (Andrews 1986) – which obviates the point of using a consistent estimator in the test.

When we enforce consistency

Recall that Z is defined as follows:

$$Z = \frac{\text{estimator of the mean} - \text{hypothesized mean}}{\text{standard deviation}/\sqrt{\text{sample size}}} = \frac{X_n - \theta_0}{\sigma/\sqrt{n}}$$

Consistency (recall): $\forall \epsilon > 0, \lim_{n \to \infty} Pr(|X_n - \theta| > \epsilon) = 0.$

When we enforce consistency

Recall that Z is defined as follows:

$$Z = \frac{\text{estimator of the mean} - \text{hypothesized mean}}{\text{standard deviation}/\sqrt{\text{sample size}}} = \frac{X_n - \theta_0}{\sigma/\sqrt{n}}$$

Consistency (recall): $\forall \epsilon > 0, \lim_{n\to\infty} Pr(|X_n - \theta| > \epsilon) = 0.$

If the hypothesis is **true** $(\theta_0 = \theta)$: $\forall z_\alpha > 0, \lim_{n\to\infty} Pr(|Z| \leqslant z_\alpha) = 1$

When we enforce consistency

Recall that Z is defined as follows:

$$Z = \frac{\text{estimator of the mean} - \text{hypothesized mean}}{\text{standard deviation}/\sqrt{\text{sample size}}} = \frac{X_n - \theta_0}{\sigma/\sqrt{n}}$$

Consistency (recall): $\forall \epsilon > 0, \lim_{n\to\infty} Pr(|X_n - \theta| > \epsilon) = 0.$

If the hypothesis is **true** $(\theta_0 = \theta)$: $\forall z_\alpha > 0, \lim_{n\to\infty} Pr(|Z| \leqslant z_\alpha) = 1$

If the hypothesis is **false** $(\theta_0 \neq \theta)$: $\forall z_\alpha > 0, \lim_{n\to\infty} Pr(|Z| \leqslant z_\alpha) = 0$

So what?

Different choices of how to prioritize errors w.r.t. either

1. loss functions / estimators
2. acceptance levels / critical values

converge on the same conclusion / decision *in the limit*.

So what?

Different choices of how to prioritize errors w.r.t. either

1. loss functions / estimators
2. acceptance levels / critical values

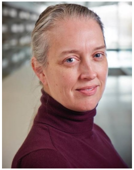
converge on the same conclusion / decision *in the limit*.

When we focus on specific cases rather than on the full range of consistent estimators, we can be more specific.

E.g., when i.i.d. sampling from a normal distribution, the sample variance s^2 converges on the population variance σ^2 as $n \to 30$ for most practical purposes.

What have we learned about inductive risk?

The modern argument



"I will argue that non-epistemic values are a required part of the internal aspects of scientific reasoning for cases where inductive risk includes risk of non-epistemic consequences." (Douglas 2000, 559)

It's inaccurate to describe Douglas's *argument* as a "revival, reiteration, or rediscovery" of Rudner's (Havstad 2022, 309).

Nevertheless...

Rudner's *example* of balancing false positives and negatives continues of play an extremely important role in the literature.

Particularly in motivating the claim that values can *legitimately* influence scientific reasoning in other contexts / domains.

We see this not just of Douglas (2000) but also in Biddle (2013), Brown (2013), Elliott (2022), Frank (2019), John (2015), Parker (2014), Plutynski (2017), Steel (2010), Steele (2013), Stegenga (2017), and Wilholt (2009).

And yet!

These statistical examples have the special feature that the influence of values will eventually wash out.

In other words:

1. Values can only influence conclusions in short/medium term
2. Conclusions should (with probability) converge over time (regardless of actually *reaching* the infinite limit)

And yet!

These statistical examples have the special feature that the influence of values will eventually wash out.

In other words:

1. Values can only influence conclusions in short/medium term
2. Conclusions should (with probability) converge over time (regardless of actually *reaching* the infinite limit)

It is an open question whether any of the extensions mentioned on the last page have similar features.

Consequences

Insofar as arguments for extending legitimate values-influence depend on the analogy to the Rudner example, those arguments bear re-examination.

More positively, our discussion suggests at least one path forward on "the new demarcation question" (Holman and Wilholt 2022):

Namely, value-influence is legitimate when? if? only if? they wash out in a manner analogous to what we find in the Rudner example.

The end

Thank you!!

Andrews, Donald W. K. (1986). Complete Consistency: A Testing Analogue of Estimator Consistency. *The Review of Economic Studies* 53.2: 263–69. DOI: 10.2307/2297650.

Annan, James D. and Julia C. Hargreaves (2011). Understanding the CMIP3 Model Ensemble. *Journal of Climate* 24: 4529–38.

Biddle, Justin B. (2013). State of the Field: Transient Underdetermination and Values in Science. *Studies in History and Philosophy of Science Part A* 44.1: 124–33. DOI: 10.1016/j.shpsa.2012.09.003.

Brown, Matt J. (2013). Values in Science Beyond Underdetermination and Inductive Risk. *Philosophy of Science* 80.5: 829–39.

Dethier, Corey (2022). When is an Ensemble Like a Sample? 'Model-Based' Inferences in Climate Modeling. *Synthese* 200.52: 1–20. DOI: 10.1007/s11229-022-03477-5.

Douglas, Heather (2000). Inductive Risk and Values in Science. *Philosophy of Science* 67.4: 559–79. DOI: 10.1086/392855.

Elliott, Kevin C. (2022). *Values in Science*. Cambridge: Cambridge University Press.

Frank, David M. (2019). Ethics of the Scientist qua Policy Advisor: Inductive Risk, Uncertainty, and Catastrophe in Climate Economics. *Synthese* 196: 3123–38. DOI: 10.1007/s11229-017-1617-3.

Havstad, Joyce C. (2022). Sensational Science, Archaic Hominin Genetics, and Amplified Inductive Risk. *Canadian Journal of Philosophy* 52.3: 295–320. DOI: 10.1017/can.2021.15.

Holman, Bennett and Torsten Wilholt (2022). The New Demarcation Problem. *Studies in History and Philosophy of Science Part A* 91: 211–20. DOI: 10.1016/j.shpsa.2021.11.011.

Jeffrey, Richard (1956). Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science* 23.3: 237–46. DOI: 10.1086/287489.

John, Stephen (2015). Inductive Risk and the Contexts of Communication. *Synthese* 192.1: 79–96. DOI: 10.1007/s11229-014-0554-7.

Parker, Wendy S. (2014). Values and Uncertainties in Climate Prediction, Revisited. *Studies in History and Philosophy of Science Part A* 46: 24–30.

Plutynski, Anya (2017). Safe or Sorry? Cancer Screening and Inductive Risk. In: *Exploring Inductive Risk: Case Studies of Values in Science*. Ed. by Kevin C. Elliott and Ted Richards. Oxford: Oxford University Press: 149–70.

Rudner, Richard (1953). The Scientist qua Scientist makes Value Judgments. *Philosophy of Science* 20.1: 1–6. DOI: 10.1086/287231.

Steel, Daniel (2010). Epistemic Values and the Argument from Inductive Risk. *Philosophy of Science* 77.1: 14–34. DOI: 10.1086/650206.

Steele, Katie (2013). The Scientist qua Policy Advisor makes Value Judgments. *Philosophy of Science* 79.5: 893–904. DOI: 10.1086/667842.

Stegenga, Jacob (2017). Drug Regulation and the Inductive Risk Calculus. In: *Exploring Inductive Risk: Case Studies of Values in Science*. Ed. by Kevin C. Elliott and Ted Richards. Oxford: Oxford University Press: 17–36.

Tokarska, Katarzyna B. et al. (2020). Past Warming Trend Constrains Future Warming in CMIP6 Models. *Science Advances* 6.12: 1–13.

Wilholt, Torsten (2009). Bias and Values in Scientific Research. *Studies in History and Philosophy of Science Part A* 40.1: 92–101. DOI: 10.1016/j.shpsa.2008.12.005.