# What Makes Statistics Valuable?

Corey Dethier

Minnesota Center for Philosophy of Science
University of Minnesota
corey.dethier@gmail.com

Nov. 5, 2023

Slides available at: https://coreydethier.com/Slides/WMSV.pdf

# What is the true value of ECS?

| Model | ECS (°C) |
|---:|:---:|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

– from Tokarska et al. (2020)

# What is the true value of ECS?

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Mean:** 3.15°C

– from Tokarska et al. (2020)

# What is the true value of ECS?

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Mean:** 3.15°C

**Median:** 2.97°C

– from Tokarska et al. (2020)

# What is the true value of ECS?

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Mean:** 3.15°C

**Median:** 2.97°C

**Spread:** 1.92 - 4.28°C

– from Tokarska et al. (2020)

# What is the point of statistics?

What is the point of statistics?

# What is the point of statistics?

What is the point of statistics?

"the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole." (Fisher 1922, 311)

## Why these tools?

Why use the methods of statistics to analyze data (as opposed to other methods)?

The methods of statistics are **epistemically efficient**: reliably discriminating and low epistemic cost.

# The plan

1. Discriminating & non-discriminating methods
2. Statistics & epistemic efficiency
3. Philosophical perspectives

# Discriminating and non-discriminating methods

# Our data

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Mean:** 3.15°C

**Median:** 2.97°C

**Spread:** 1.92 - 4.28°C

– from Tokarska et al. (2020)

# Methods of analyzing data

**Method**: a (bi)conditional that says which hypothesis to prefer.

- Prefer $h$ iff $h =$ mean of the sample.
- Prefer $h$ iff $h =$ the spread of the sample.

These examples are **perfectly discriminating**: they always tell us to prefer exactly one hypothesis.

# Example 1: Agreement

**Agreement:** prefer $h$ iff all of the estimates entail that $h$ is true.

# Example 1: Agreement

**Agreement:** prefer $h$ iff all of the estimates entail that $h$ is true.

Agreement is **negatively discriminating**: it always recommends preferring at least one hypothesis.

# Example 1: Agreement

**Agreement:** prefer *h* iff all of the estimates entail that *h* is true.

Agreement is **negatively discriminating**: it always recommends preferring at least one hypothesis.

Agreement is not (always) **positively discriminating**: in some circumstances, it recommends preferring multiple hypotheses.

# Agreement: a good case

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Hypothesis space:**

- $h_1$ : ECS $< 1.5°$C
- $h_2$ : $1.5°$C $\leqslant$ ECS & ECS $< 4.5°$C
- $h_3$ : $4.5°$C $\leqslant$ ECS

# Agreement: a good case

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Hypothesis space:**

- $h_1$ : ECS $< 1.5°$C
- $h_2$ : $1.5°$C $\leqslant$ ECS & ECS $< 4.5°$C
- $h_3$ : $4.5°$C $\leqslant$ ECS

Agreement: prefer $h_2$.

# Agreement: a bad case

| Model | ECS (°C) |
|---:|:---:|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Hypothesis space:**

- $h_1 : 1.5°C \leqslant \text{ECS}$
  $\& \ \text{ECS} < 4.5°C$
- $h_2 : 1.5°C \leqslant \text{ECS}$
  $\& \ \text{ECS} < 5.5°C$

# Agreement: a bad case

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

**Hypothesis space:**

- $h_1 : 1.5°\mathrm{C} \leqslant \mathrm{ECS}$ $\ \& \ \mathrm{ECS} < 4.5°\mathrm{C}$
- $h_2 : 1.5°\mathrm{C} \leqslant \mathrm{ECS}$ $\ \& \ \mathrm{ECS} < 5.5°\mathrm{C}$

Agreement: prefer $h_1$ & $h_2$.

# Example 2: Consensus

Suppose that every estimate has an expected error of $\pm 2^\circ$C.

**Consensus:** prefer $h$ iff $h$ includes all and only the values that fall within $\pm 2^\circ$C of every estimate.

# Example 2: Consensus

Suppose that every estimate has an expected error of $\pm 2^{\circ}$C.

**Consensus:** prefer $h$ iff $h$ includes all and only the values that fall within $\pm 2^{\circ}$C of every estimate.

Consensus is not (always) **negatively discriminating**: it doesn't always recommends preferring at least one hypothesis.

Agreement is **positively discriminating**: it always recommends preferring at most one hypothesis.

# Consensus: a good case

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

Consensus: prefer $h$:

ECS is in 2.28 - 3.92°C

## What distinguishes consensus from agreement?

Consensus makes use of **higher-order evidence**: evidence about the (expected) accuracy of the individual estimates.

# What distinguishes consensus from agreement?

Consensus makes use of **higher-order evidence**: evidence about the (expected) accuracy of the individual estimates.

Consensus is **reliably discriminating**: if your assumptions are correct, its recommendations are trustworthy.

# What distinguishes consensus from agreement?

Consensus makes use of **higher-order evidence**: evidence about the (expected) accuracy of the individual estimates.

Consensus is **reliably discriminating**: if your assumptions are correct, its recommendations are trustworthy.

Consensus is **costly**: in order to use consensus, you need (accurate) higher-order evidence.

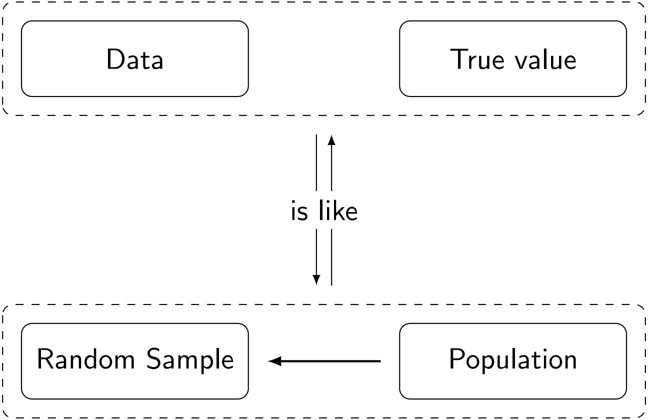
# Statistics and epistemic efficiency

# What if we want to use statistics?

| Model | ECS (°C) |
|---|---|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

– from Tokarska et al. (2020)

# Statistical reasoning



# Statistical reasoning

Data

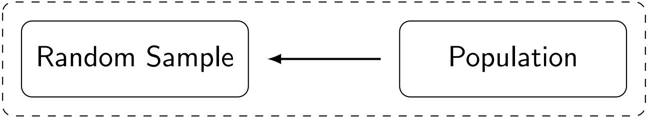
True value

1. Re-describe the data as a sample
2. Infer the population from the sample
3. Infer the true value from the population

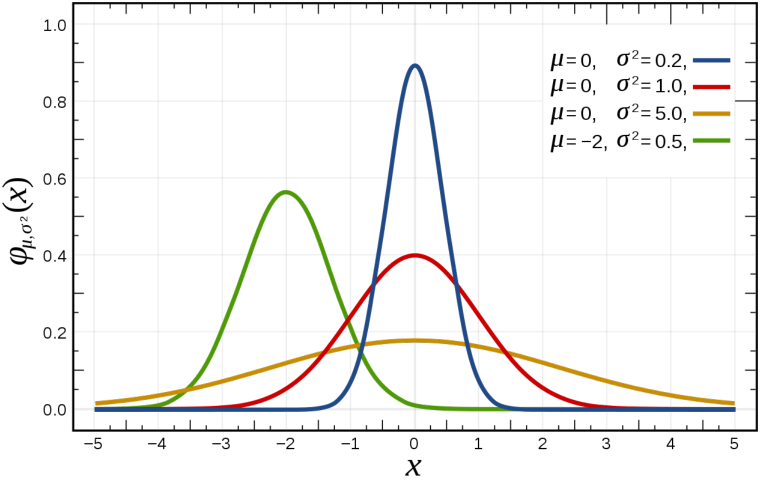
Random Sample $\longleftarrow$ Population

# Statistical reasoning


1. Re-describe the data as a sample
2. Infer the population from the sample
3. Infer the true value from the population



# Re-describing the data, pt. 1




– Wikimedia Commons

## Re-describing the data, pt. 2

Re-describe the data as a **probability density function** $f(x)$:

"Center" $\rightarrow$ 1st moment (mean): $\int xf(x)dx$.

"Width" $\rightarrow$ 2nd central moment (variance): $\int (x-\mu)^2 f(x)dx$.

"Irregularities" $\rightarrow$ higher standardized moments.

3rd standardized moment (skewness): $\int (\frac{x-\mu}{\sigma})^3 f(x)dx$

4th standardized moment (kurtosis): $\int (\frac{x-\mu}{\sigma})^4 f(x)dx$

# Re-describing the data, part 3

| Model | ECS (°C) |
|---:|:---:|
| BCC-ESM1 | 3.23 |
| CNRM-CM6-1-HR | 4.28 |
| EC-Earth3 | 4.20 |
| GDFL-CM4 | 3.87 |
| GDFL-ESM4 | 2.62 |
| GISS-E2-1-G | 2.72 |
| INM-CM5-0 | 1.92 |
| MICRO6 | 2.57 |
| MPI-ESM1-2-HR | 2.97 |
| NorESM2-LM | 2.60 |
| SAM0-UNICON | 3.72 |

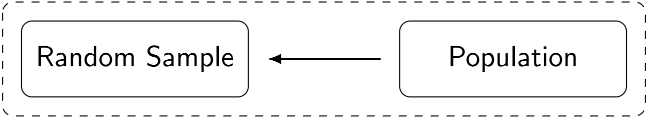
**S. mean** ($\bar{x}$): 3.15°C

**S. variance** ($s^2$): .61°C

– from Tokarska et al. (2020)

# We're more than halfway there


1. Re-describe the data as a sample
2. Infer the population from the sample
3. Infer the true value from the population



# From sample to population, part 1

To continue, we need a **statistical model**:

1. A specification of population “family” (normal)
2. A specification of sampling procedure (random / IID)

Given this statistical model, the relationship between the sample and population is given by the *t*-distribution.

# From sample to population, part 2

We calculate "$t$-scores" for each hypothesis concerning $\mu$:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{3.15 - \mu}{\frac{.78}{\sqrt{11}}}$$

The probability of $t$-scores is given by the $t$-distribution:

$$p(x < t < y) = \int_x^y \frac{(\frac{n}{2} - 1)!}{(\frac{n-1}{2} - 1)!\sqrt{(n-1)\pi}} \left(1 + \frac{t^2}{n-1}\right)^{\frac{-n}{2}} dt$$

# From sample to population, part 3

The "critical values" for the mass of the $t$-distribution for $n = 11$:

| | .5 | .8 | .95 | .98 | .99 | .995 |
|---|---|---|---|---|---|---|
| $c$ | .70 | 1.37 | 1.81 | 2.23 | 2.76 | 3.17 |

Which means, e.g.,

$$p(-1.37 < t < 1.37) = .8$$

For a given level of confidence, you get exactly one preferred hypothesis. E.g: for .95, prefer $h$: ECS is between 2.72 - 3.58$^\circ$C.

# The intuition

# The intuition

# The intuition

# The intuition

Each new observation provides us with:

1. an estimate of the true value of $\mu$;
2. evidence about the accuracy of the other estimates.

The mean combines all the estimates; the higher moments (e.g., $s^2$) combine the evidence about their accuracy.

# The role of higher-order evidence

Like Consensus, statistical methods rely on higher-order evidence.

But where Consensus rely the expected accuracy of the individual measurements, statistical methods rely on the expected accuracy of the distribution *as a whole*.

Which methods should we prefer?

# The role of higher-order evidence

Like Consensus, statistical methods rely on higher-order evidence.

But where Consensus rely the expected accuracy of the individual measurements, statistical methods rely on the expected accuracy of the distribution *as a whole*.

Which methods should we prefer? Depends on which kind of information we have (or can get).

# The role of higher-order evidence

Like Consensus, statistical methods rely on higher-order evidence.

But where Consensus rely the expected accuracy of the individual measurements, statistical methods rely on the expected accuracy of the distribution *as a whole*.

Which methods should we prefer? Depends on which kind of information we have (or can get).

But not a satisfying answer to “why these methods?”

# The final step

Data | True value

1. Re-describe the data as a sample
2. Infer the population from the sample
3. Infer the true value from the population

Random Sample ← Population

# Recall our assumptions

The inference from sample to population relied on two assumptions:

1. A specification of population “family” (normal)
2. A specification of sampling procedure (random / IID)

# Recall our assumptions

The inference from sample to population relied on two assumptions:

1. A specification of population "family" (normal)
2. A specification of sampling procedure (random / IID)

The received view in statistics is that "all models are wrong" (Box 1979). Our assumptions are—at best!—worthwhile idealizations or approximations.

# The last step is substantive!

"If the statistician thoughtlessly decides, whatever be the test, to reject an hypothesis when $P \leqslant .01$, say, and accept it when $P > .01$, it will make a considerable difference to his conclusions whether he uses [one test statistic or another]. But as the ultimate value of statistical judgment depends on a clear understanding of the meaning of the statistical tests applied, the difference between the values of the two $P$'s should present no difficulty." (Neyman and Pearson 1928, 192; quoted in Mayo 1996, 386)

# So what does statistics do for us?

After carrying out a statistical test, we know:

> *If* (a) approximately normal and (b) approximately random sampling, *then* $h$: ECS is between 2.72 - 3.58$^{\circ}$C is preferable at the .95 level.

We're buying discriminating power; the cost is the assumptions required for the statistical model.

# We can describe other methods in the same way

With Consensus, the cost is our assumptions about the accuracy of individual estimators.

With “just accept the sample mean,” the cost is effectively an assumption that the sample mean is perfectly accurate.

# We can describe other methods in the same way

With Consensus, the cost is our assumptions about the accuracy of individual estimators.

With "just accept the sample mean," the cost is effectively an assumption that the sample mean is perfectly accurate.

Relative to the discriminatory power that they offer, the epistemic cost (and thus the literal cost) of the statistical assumptions is relatively low.

# Three perspectives on the cost

# Philosophical approaches to statistics

"Statistical methods" are extremely varied, and while they all require some sort of assumptions (some sort of "statistical model"), the exact nature of the assumptions differs dramatically.

We can view disagreements in the philosophy of statistics as grounded in disagreements about the "cost" of different methods.

# A classical perspective

Classical statisticians primarily think about cost in terms of experimental control.

"We were certainly aware that inferences must make use of prior information ... [but] we came to the conclusion, rightly or wrongly, that **it was so rarely possible to give sure numerical values to these entities, that our line of approach must proceed otherwise.** Thus we came down on the side of using only those probability measures that could be related to relative frequency." (Pearson 1962, 395–96)

See also Fisher (1973, 37), Neyman (1952, 22–27), or Mayo (1996).

# Classical answer to the question

Why these methods?

Because they reliably discriminate between hypotheses while requiring little more than what we can experimentally control.

# A personalist perspective

For a strict personalist, the "cost" of the assumptions is already built into your prior distribution.

What you want is a method that doesn't add any additional cost—that delivers results that are logically determined by the assumptions you started with.

See, e.g., Howson and Urbach (2006, 301).

# Personalist answer to the question

Why these methods?

Because they reliably discriminate between hypotheses in a "logical" way.

# The instrumental perspective

Many practitioners—Cox (2006), Gelman and Shalizi (2013), and Kass (2011)—adopt a much more instrumental perspective.

They reject the view (held by both parties) that subjective priors are largely uncontrollable.

Instead, priors are just like any other part of the model.

# The efficiency answer

Why these methods?

Because they discriminate *efficiently*: they reliably discriminate while requiring a relatively low cost.

But what counts as a low epistemic cost is (of course) context-sensitive!

What makes statistics valuable? Its diversity: it offers tools that are efficient in a wide variety of situations.

Box, George E. P. (1979). Robustness in the Strategy of Scientific Model Building. In: *Robustness in Statistics*. Ed. by Robert L. Launer and Graham N. Wilkinson. New York: Academic Press: 201–36.

Cox, David R. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

Fisher, Ronald A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A* 222.594-604: 309–68. DOI: 10.1098/rsta.1922.0009.

— (1973). *Statistical Methods and Scientific Inference*. 3rd ed. New York: Macmillan.

Gelman, Andrew and Cosma Rohilla Shalizi (2013). Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology* 66: 8–38. DOI: 10.1111/j.2044-8317.2011.02037.x.

- Howson, Colin and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach*. 3rd ed. Chicago: Open Court.
- Kass, Robert E. (2011). Statistical Inference: The Big Picture. *Statistical Science* 26.1: 1–9. DOI: 10.1214/10-STS337.
- Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.
- Neyman, Jerzy (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. 2nd ed. Washington, D.C.: US Department of Agriculture.
- Neyman, Jerzy and Egon S. Pearson (1928). On the Use and Interpretation of Certain Test Criteria for the Purposes of Statistical Inference. *Biometrika* 20A: 175–240, 263–94. DOI: 10.2307/2331945.
- Pearson, Egon S. (1962). Some Thoughts on Statistical Inference. *The Annals of Mathematical Statistics* 33.2: 394–403. DOI: 10.1214/aoms/1177704566.

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