Department of Philosophy, Logic and Scientific Method
London School of Economics

What constitutes evidence and its role in calibration and confirmation

Dr Charlotte Werndl, Associate Professor (c.s.werndl@lse.ac.uk)

An Example from Chemical Engineering

- Biochemical oxygen demand is the amount of dissolved oxygen needed by aerobic biological organisms to break down organic material present in a water sample over a specific time period.
- Consider a model of biological oxygen demand y as a function of time x:

$$y = k_1[1 - \exp(-k_2 x)], \tag{1}$$

where k_1 is the deoxygenation rate constant and k_2 is the reaction rate constant.

An Example from Chemical Engineering

- Biochemical oxygen demand is the amount of dissolved oxygen needed by aerobic biological organisms to break down organic material present in a water sample over a specific time period.
- Consider a model of biological oxygen demand y as a function of time x:

$$y = k_1[1 - \exp(-k_2 x)], \tag{1}$$

  where k_1 is the deoxygenation rate constant and k_2 is the reaction rate constant.
- Suppose that k_1 and k_2 are unknown and estimated from data about the biochemical oxygen demand y and time points x (calibration).

The Problem

- Can one then use the same data to confirm the model?
- This would be double-counting.
- This issue arises in all the sciences and is often hotly debated.

A Popular Position

- Many scientists endorse the position that the same data cannot be used for calibration and confirmation.
- *"If the model has been tuned to give a good representation of a particular observed quantity, the agreement with that observation cannot be used to build confidence in that model."* (IPCC report)
- Many philosophers, e.g., Worrall (2002, 2008), endorse a similar position.

A Popular Position

- Many scientists endorse the position that the same data cannot be used for calibration and confirmation.
- *"If the model has been tuned to give a good representation of a particular observed quantity, the agreement with that observation cannot be used to build confidence in that model."* (IPCC report)
- Many philosophers, e.g., Worrall (2002, 2008), endorse a similar position.
- Against these positions, it is argued that double-counting is legitimate.

- Introduction
- Comparative Confirmation
- Non-Comparative Confirmation
- Inductive Problems
- Concluding Remarks

- Introduction
- Comparative Confirmation
- Non-Comparative Confirmation
- Inductive Problems
- Concluding Remarks

Probabilistic Confirmation Theory

- Use probabilistic confirmation theory to tackle question about double-counting. I.e.:

$$Pr(\text{Model}|\text{Evidence}) = \frac{Pr(\text{Evidence}|\text{Model})Pr(\text{Model})}{Pr(\text{Evidence})}.$$

- Start with case where performance of two specific hypotheses is compared (comparative confirmation).

A Simple Example

Let us start with a very simple case. Base models:

- M: $y(t) = mt$
- N: $y(t) = nt^2$

Model instances M_1, ..., $N1$, ... assign particular values to free parameters m, n.

The Question of Double-Counting in this Framework

- If data is used to determine which instance of a base hypothesis is true, can this data also serve to *confirm* (i.e. raise the probability of) the base hypothesis?

A Simple Example

So much for the models, but what hypotheses are we interested in?

It depends on how the models are perceived vis-à-vis the real world.

- Are the models supposed to be exact replicas of (some aspect) of the climate system?
- Or are the observations or models known to be imperfect?

There are a variety of cases; here I just consider one (observational error).

A Simple Example

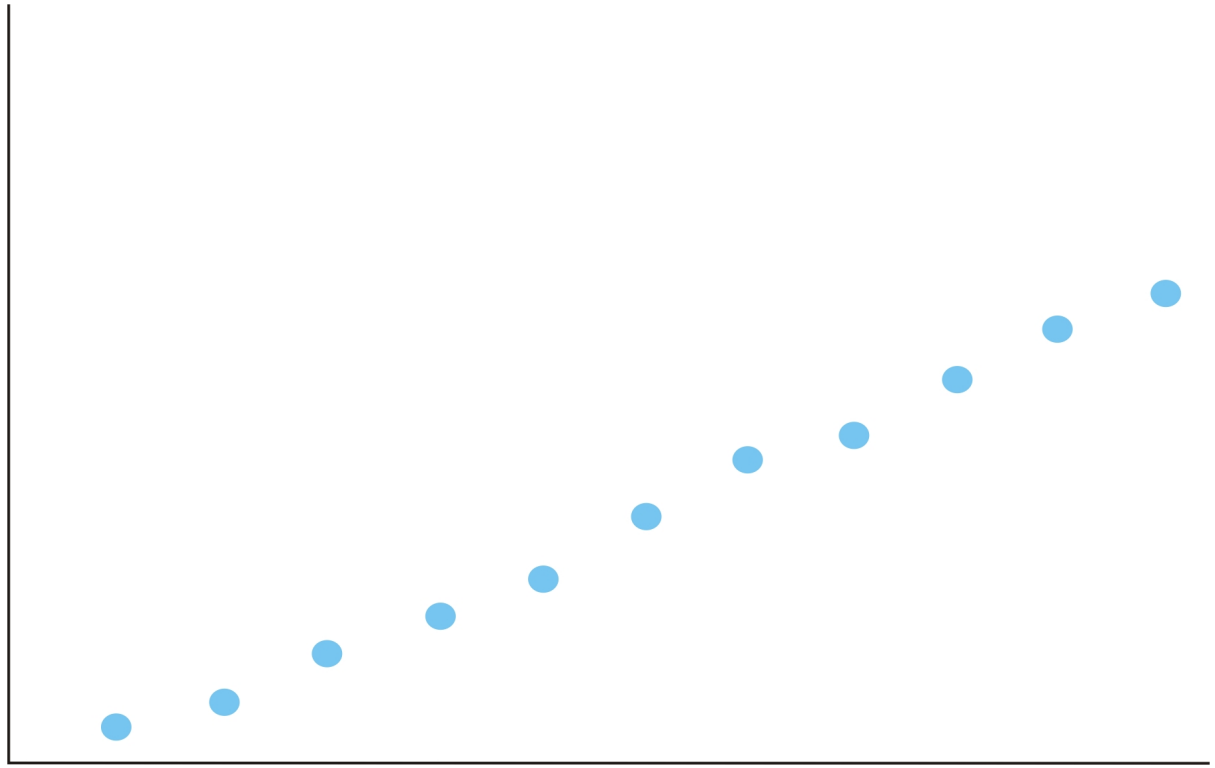
There may be observational error. In simple case can be modeled by, e.g., a Gaussian distribution.

Re-specify models to include error:

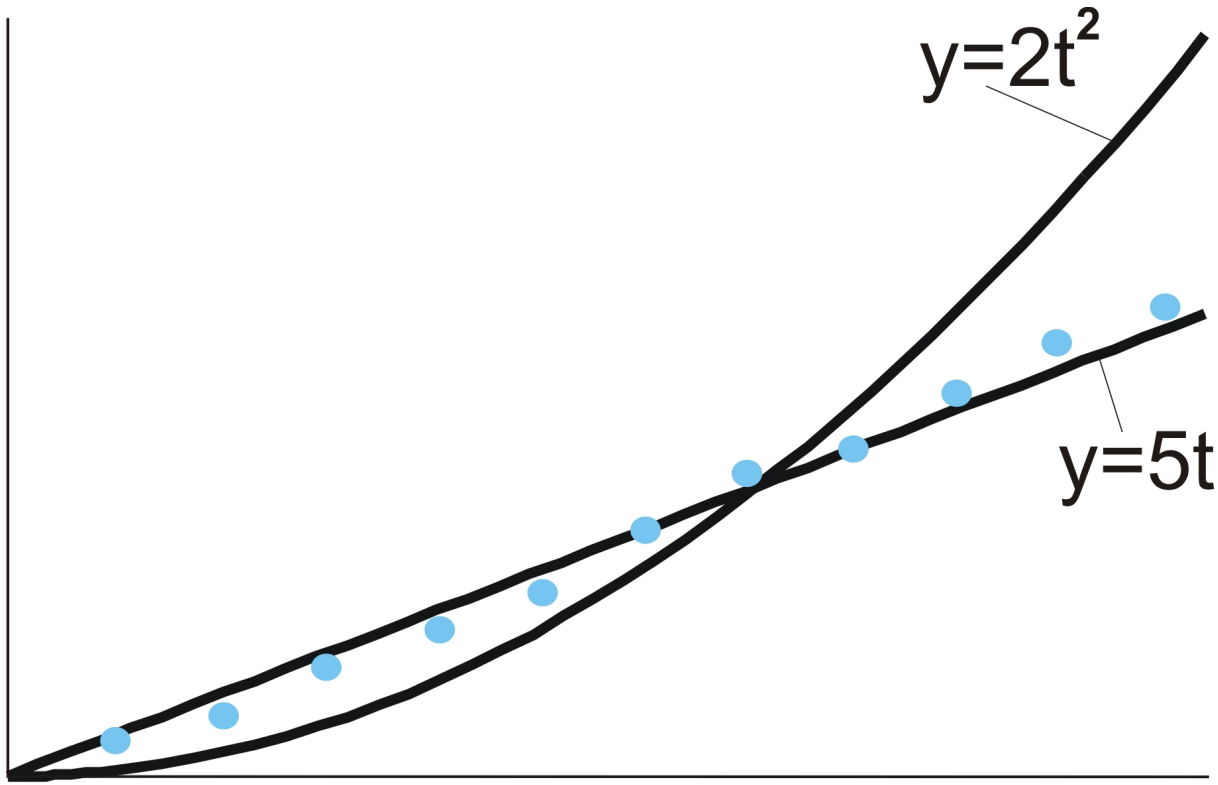
- $M : y(t) \sim mt + N(0, \sigma)$
- $N : y(t) \sim nt^2 + N(0, \sigma)$

Here M_1 denotes 'Model M with parameter values labelled '1' describes generation of $y(t)$'.

Eleven data points



A Simple Example



The following model instances provide the best fit to the data:

M_5: $y(t) \sim 5t + N(0, \sigma)$; N_2: $y(t) \sim 2t^2 + N(0, \sigma)$.

A Simple Example

- M_5 has a much better fit with the data than N_2. That is, $Pr(E|N_2) < Pr(E|M_5)$.

Then M is confirmed relative to N.

A Simple Example

- The eleven data points are legitimately used for both calibration and confirmation.

A Simple Example

- The eleven data points are legitimately used for both calibration and confirmation.
- In general: the Bayesian analysis shows that M can be confirmed relative to N because one model has a better 'fit' with data than the other. That is, the likelihoods $Pr(E|M_i)$ and $Pr(E|N_j)$ differ.
- Here concerns about double-counting of scientists and philosophers (e.g., Worrall 2002, 2008) are misplaced.

- Introduction
- Comparative Confirmation
- Non-Comparative Confirmation
- Inductive Problems
- Concluding Remarks

Non-Comparative Confirmation

- This is a matter of whether the evidence E confirms a base model M tout court, i.e. relative to its full complement $\neg M$.
- As for comparative confirmation: M can be confirmed tout court, even when there is calibration, and worries about double-counting are misplaced.

Example of Biochemical Oxygen Demand

- Consider again the model of biological oxygen demand y as a function of time x:

$$y = k_1[1 - \exp(-k_2 x)], \tag{2}$$

  where k_1 is the deoxygenation rate constant and k_2 is the reaction rate constant.

- If data are used to estimate k_1 and k_2 (calibration) the same data can also confirm the model.

Inductive Problems

- When scientists debate the legitimacy of double-counting, their focus is on the wrong problem.
- Behind these debates there are three other problems.
- These problems do not show that double-counting is illegitimate, but that the confirmation and the inductive reasoning might fail.

Inductive Problem 1: Good Fit With Any Data

- Suppose that, whatever the data, there will be a good fit with the model M.
- E.g., polynomial model with 100 free parameters will provide a good fit to any arbitrary 100 data.

Inductive Problem 1: Good Fit With Any Data

- Then scientists often think that both M and $\neg M$ are equally successful, i.e., $P(E|M) = P(E|\neg M)$
  (those hypotheses in $\neg M$ that do better than M are counter-balanced by those hypotheses in $\neg M$ that do worse than M).
- Then: there is calibration but no confirmation.

Inductive Problem 1: Good Fit With Any Data

- Then scientists often think that both M and $\neg M$ are equally successful, i.e., $P(E|M) = P(E|\neg M)$
  (those hypotheses in $\neg M$ that do better than M are counter-balanced by those hypotheses in $\neg M$ that do worse than M).
- Then: there is calibration but no confirmation.
- This concerns the failure rather than the illegitimacy of confirmation/double-counting.

Inductive Problem 2: Relevant Evidence

It may be disputable whether the evidence is relevant. For instance, the worry may be that:

- The lifespan of the model is the medium-run future, and evidence concerns only the past.
- Underlying thought: the model does not include the main processes relevant for the medium-run future *and* the past.

Inductive Problem 2: Relevant Evidence

- If lifespan of model is medium-run future: past data cannot be used for calibration/confirmation.

E.g., climate scientists raise this worry:

> *Statements about future climate relate to a never before experienced state of the system; thus, it is impossible to either calibrate the model for the forecast regime of interest or confirm the usefulness of the forecasting process (Stainforth et al. 2007a, 2146).*

Inductive Problem 2: Relevant Evidence

- If lifespan of model is medium-run future: past data cannot be used for calibration/confirmation.

E.g., climate scientists raise this worry:

> *Statements about future climate relate to a never before experienced state of the system; thus, it is impossible to either calibrate the model for the forecast regime of interest or confirm the usefulness of the forecasting process (Stainforth et al. 2007a, 2146).*

- Again: issue here is the *failure*, rather than the *legitimacy*, of calibration/confirmation/double-counting.

Inductive Problem 3: Radical Uncertainty

When the relevant processes are very poorly understood. . .

- . . . there is, plausibly, much uncertainty about the other possible models $\neg M$.
- In such case we are unable to assess even roughly how likely the evidence is given the other possible models, and so there can be no confirmation.

Inductive Problem 3: Radical Uncertainty

- Some climate scientists indeed seem to argue that non-comparative confirmation and thus double-counting *fails* due to this radical uncertainty.

  We take climate ensembles exploring model uncertainty as potentially providing a lower bound on the maximum range of uncertainty and thus a non-discountable [unable-to-be-ignored] climate change envelope [range of climate-change predictions]. (Stainforth et al. 2007b, 2167)

Inductive Problem 3: Radical Uncertainty

- Some climate scientists indeed seem to argue that non-comparative confirmation and thus double-counting *fails* due to this radical uncertainty.

  We take climate ensembles exploring model uncertainty as potentially providing a lower bound on the maximum range of uncertainty and thus a non-discountable [unable-to-be-ignored] climate change envelope [range of climate-change predictions]. (Stainforth et al. 2007b, 2167)

- Again: the issue concerns the *failure*, rather than the *legitimacy*, of confirmation/double-counting.

Induction and Inductive Reasoning

- Analysis of failures of induction with the aim of understanding induction has a long tradition in philosophy (e.g. Carnap, Hume).
- Analysis of the three problems (relevant evidence, radical uncertainty, good fit with any data) contribute to this.

- Introduction
- Comparative Confirmation
- Non-Comparative Confirmation
- Inductive Problems
- Concluding Remarks

Concluding Remarks

- More clarity needed in science literature.
- I argued against common view that separate data are needed for calibration and confirmation. Double-counting is legitimate.
- Scientists' worries most charitably reconstructed as concerns about induction: confirmation/double-counting might fail because
  - Model has good fit with any arbitrary data;
  - Past evidence is irrelevant for assessing models that concern the medium-run future;
  - Radical uncertainty how well other models could explain the data.

Steele, K. and Werndl, C. (2013). 'Climate Models, Confirmation and Calibration'. *The British Journal for the Philosophy of Science* 64, 609-635.