Target trial emulation and Bayesian generative models

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

Leuven, September 2024

DEPARTMENT OF
STATISTICS

Target trial emulation

Causal analysis of observational data is hard. This is not just because we require assumptions that lie outside of the data (e.g. SUTVA and unconfoundedness). Even if these assumptions hold, we must be careful not to introduce biases into the analysis by using the data inappropriately.

Recently, Hernan and Robins (2016) described the use of target trial emulation as a qualitative approach to ensure rigour in the causal analysis of observational data.

The idea of conceptualising an idealized randomized trial when designing observational studies dates back a long time. The recent innovation is in retrospective analysis of observation studies (that already exist) and the notio of a closest matching RCT

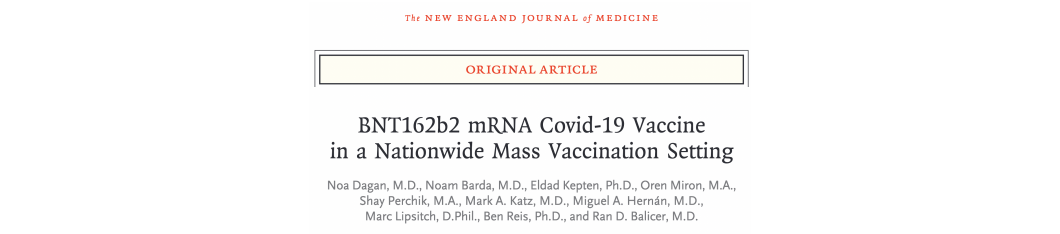
Target trial emulation

In Target Trial Emulation, the analyst is required to frame their causal questions by specifying the protocol of an explicit pragmatic (target) randomized trial, i.e. design a matched target trial to the observational study, that they would have liked to carry out.

The analyst then follows the protocol of the target trial in the analysis and reporting of inference from the observational data.

The operational conduct of randomized controlled trials (RCTs) require pre-specification of the trial protocol, including features such as eligibility criteria for the trial and treatments. Carrying this over to causal analysis of observational studies improves the rigour and quality of the analysis

Example: effectiveness of the Pfizer-BioNTech vaccine



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

BNT162b2 mRNA Covid-19 Vaccine
in a Nationwide Mass Vaccination Setting

Noa Dagan, M.D., Noam Barda, M.D., Eldad Kepten, Ph.D., Oren Miron, M.A., Shay Perchik, M.A., Mark A. Katz, M.D., Miguel A. Hernán, M.D., Marc Lipsitch, D.Phil., Ben Reis, Ph.D., and Ran D. Balicer, M.D.

Target trial emulation was used to investigate the efficacy of the Pfizer-BioNTech COVID-19 vaccine using data from Israel's largest health care organisation.

The large sample size ($n \approx 600,000$) allowed the researchers to estimate vaccine effectiveness in subpopulations that previous randomized trials could not evaluate due to lack of power, e.g. individuals over 70.

Target trial protocol specification

Protocol component	Protocol specification
Eligibility criteria	Individuals aged ≥ 16 without a previous vaccination and documented positive PCR test, and has been a member of the healthcare organisation for the previous 12 months.
Treatment strategies	1. One dose of the vaccine at baseline and one dose 3 weeks later 2. No vaccination
Assignment	Individuals are randomly assigned to either strategy at baseline and are aware of their assigned strategy, i.e. no double-blind assignment.
Time zero & follow-up	Starts at assignment and ends at diagnosis of COVID-19 outcome, death, loss to follow-up, or administrative end of follow-up.
Outcomes	COVID-19 diagnosis, COVID-19 hospitalization, severe COVID-19 outcome, COVID-19 death.
Causal estimands	Per-protocol effect

First, the user must frame their causal questions by carefully specifying the protocol of a pragmatic randomized trial that they would have liked to carry out.

Target trial protocol emulation

Protocol component	Protocol emulation
Eligibility criteria	Same as for target protocol
Treatment strategies	Same as for target protocol
Assignment	Individuals are assumed to be randomly assigned to either strategy conditional on a set of measured confounders, e.g. age, sex, area of residence, prior infection, comorbidities etc.
Time zero & follow-up	For vaccinated, eligible individuals: time zero is the day of vaccination. For unvaccinated, eligible individuals: time zero is the first day of eligibility.
Outcomes	Same as for target protocol
Causal estimands	Same as for target protocol

The analysis of the observational data must then emulate the protocol as closely as possible.

Target trial workflow

Target trial emulation provides a qualitative workflow for causal analysis

This ensures that key variables and analysis objectives are pre-sepcified, improving reproducibilty and transparency of the analysis

We highlight that generative models can be used within target trial emulation to provide a complete, self-contained, quantitative causal framework without the need for counterfactuals.

'Causal inference can be classified into two distinct classes of problems: predicting effects of interventions and reasoning about counterfactuals. - Judea Pearl

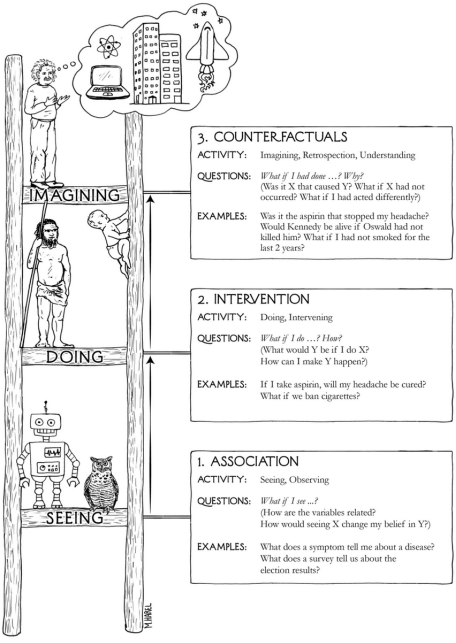
Most medical applications are of the first type, and we offer a bespoke framework that avoids any additional redundant machinery.

Pearl's causal ladder

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.



- Potential outcomes
- Target trial predictive framework
- Probability

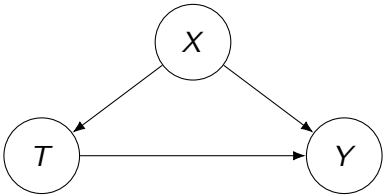
Set-up

We will consider the simplest (canonical) set-up for illustration.

We have data $Z_i = (X_i, T_i, Y_i)$ $(i = 1, \ldots, n)$, recorded on n independent units from an observational study where

- X_i is a vector of pre-treatment covariates
- Y_i is an outcome variable of interest
- T_i is a binary treatment indicator, $T_i \in \{0, 1\}$

We have potential confounding and we wish to infer something about the causal treatment effect.



The idea

- We treat causal inference as a missing data problem, where the missing data is from a closest matching, hypothetical, population-scale randomized controlled trial (RCT), matched to the observational study
- We use a predictive generative model (joint probability distribution) to then simulate participants and outcomes from the population RCT, conditional on information from the observational study
- Following which any scientific (refutable) causal quantity of interest can be read off from the synthetic RCT data, without the need to introduce counterfactuals
- Repeated simulation of the population-scale RCT quantifies uncertainty in the population causal effect

Foundations

In the paper we develop an axiomatic causal framework without the need for counterfactuals – building on decision-theoretic causal inference (e.g. Dawid 2021) and target trial emulation.

A regime indicator F_T indexes the joint distribution of observables under different regimes:

- $P(Y, T, X \mid F_T = \mathscr{O})$: observational regime
- $P(Y, T, X \mid F_T = \mathscr{E})$: closest matching experimental (randomized) regime

We'd like data $Z^{\mathscr{E}} = (Y^{\mathscr{E}}, X^{\mathscr{E}}, T^{\mathscr{E}})$ but we have $Z^{\mathscr{O}} = (Y^{\mathscr{O}}, X^{\mathscr{O}}, T^{\mathscr{O}})$.

Solution: build a joint predictive (generative) model conditioned on what you know

$$P(Z^{\mathscr{E}} \mid z^{\mathscr{O}}_{i=1:n})$$

Simulate $Z^{\mathscr{E}}_{n+1:N}$ for very large N for the missing population on the target RCT and pick off any (observable) causal quantity of interest.

Causal inference as a missing data problem

Recall that causal inference can be framed as a missing data problem in the Rubin causal model.

Unit	Y^0	Y^1	T	X
1	?	Y_1^1	1	X_1
2	Y_2^0	?	0	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	?	Y_n^1	1	X_n

Only one potential outcome per unit is ever observed, the other being counterfactual on assignment of the treatment to a unit.

Target trial prediction – data table

Unit	Y	T	X	F_T
1	Y_1	1	X_1	$\mathscr{O}$
2	Y_2	0	X_2	$\mathscr{O}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_n	1	X_n	$\mathscr{O}$
$n+1$	?	?	?	$\mathscr{E}$
$n+2$	?	?	?	$\mathscr{E}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

By assumption, once we have $Z_{n+1:N}$, we can exactly recover the target parameter $\theta(Z_{n+1:N})$, e.g. the population average treatment effect is the Neyman difference-in-means estimator

$$\theta^{ATE}(Z_{n+1:\infty}) = \lim_{N\to\infty} \left[\frac{\sum_{k=n+1}^{N} Y_k 1(T_k = 1)}{\sum_{k=n+1}^{N} 1(T_k = 1)} - \frac{\sum_{k=n+1}^{N} Y_k 1(T_k = 0)}{\sum_{k=n+1}^{N} 1(T_k = 0)} \right].$$

Prequential factorization

We require a joint predictive distribution $p(z_{n+1:N} \mid z_{1:n}, F_T = \mathscr{E})$.

Given a joint predictive we can then impute the missing data conditioned on the observational data $Z_{1:n}$.

But specifying a joint predictive generative model is challenging. To make this task more manageable, it is helpful to use the chain rule to decompose the joint into a product of conditional factors

$$p(z_{n+1:N} \mid z_{1:n}, F_T = \mathscr{E}) = \prod_{i=n+1}^{N} p(z_i \mid z_{1:i-1}, F_T = \mathscr{E}).$$

Predictive resampling

We use the shorthand $p_k(z) = p(z \mid z_{1:k}, F_T = \mathscr{E})$.

$$
\begin{aligned}
\text{Draw } Z_{n+1} \sim p_n, \text{ then update the predictive with } Z_{1:n+1} \\
\text{Draw } Z_{n+2} \mid Z_{n+1} \sim p_{n+1}, \text{ then update the predictive with } Z_{1:n+2} \\
\text{Draw } Z_{n+3} \mid Z_{n+2} \sim p_{n+2}, \text{ then update the predictive with } Z_{1:n+3} \\
\vdots
\end{aligned}
$$

This computational scheme—introduced by Fong et al. (2022)—is called predictive resampling.

Predictive causal structure

We enforce each p_k to factorize as

$$p_k(y, t, x) = p_k(y \mid t, x) p_k(t) p_k(x),$$

to reflect the dependence structure of the target trial. We can then decompose each step of the predictive resampling as follows.

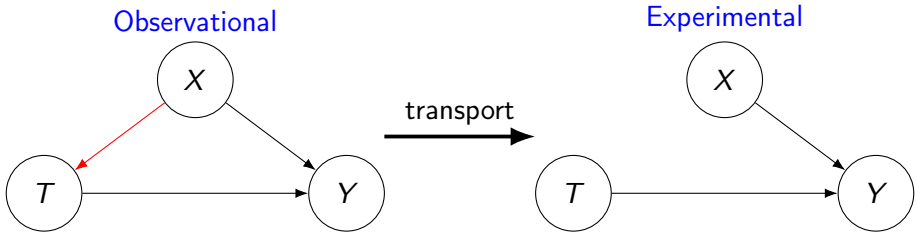
Conditional on $Z_{1:k-1}$:

1. Sample X_k from $p_{k-1}(x)$ *(Predict the pre-treatment covariates for the k-th unit)* – for example using a Bootstrap (eCDF) to draw a datum to copy
2. Sample T_k from Ber(0.5) *(Assign a randomized treatment value)*
3. Sample Y_k from $p_{k-1}(y \mid T_k, X_k)$ *(Predict the outcome given the administered treatment and pre-treatment covariates)*
4. Update the predictive to $p_k \mid X_k, T_k, Y_k$

Building a predictive from observational data

A conspicuous problem is the absence of target trial data with which to build our first-step predictive $p_n(z) = p(z \mid z_{1:n}, F_T = \mathscr{E})$.

To resolve this, we instead elicit a first-step predictive for the observational regime $p(z \mid z_{1:n}, F_T = \mathscr{O})$ and transport it into the experimental regime using an axiomatic framework.



Transportability

Suppose we have specified a first-step predictive $p(z \mid z_{1:n}, \mathscr{O})$. We will first assume that the population characteristics do not change from the observational to the experimental regimes:

$$p(x \mid z_{1:n}, \mathscr{E}) = p(x \mid z_{1:n}, \mathscr{O}).$$

Israel vaccine example: in the observational data, X includes age, sex, area of residence, prior infection etc. We would require our target population (e.g. the population of Israel) to share the same distribution on these variables. This can be relaxed if we have additional information on our target population (e.g. census data) to construct a different predictive.

Transportability

We also assume stability in the outcome regression model

$$p(y \mid z_{1:n}, t, x, \mathscr{E}) = p(y \mid z_{1:n}, t, x, \mathscr{O}).$$

Israel vaccine example: the outcome Y is COVID-19 hospitalization and the treatment T is two doses of the Pfizer-BioNTech vaccine. So we assume that if we know the individual's T and X, their probability of becoming hospitalized from COVID-19 does not depend on how the treatment was assigned, i.e. under intervention or observation.

The g-formula density

Finally, we assume that treatment allocation is fully randomized in our target trial:

$$p(t = 1 \mid z_{1:n}, \mathscr{E}) = 0.5.$$

Our first-step predictive for the experimental regime is now given by the g-formula density

$$p(y, t, x \mid z_{1:n}, \mathscr{E}) = p(y \mid z_{1:n}, t, x, \mathscr{O})p(t \mid \mathscr{E})p(x \mid z_{1:n}, \mathscr{O}).$$

The only change from $\mathscr{O}$ to $\mathscr{E}$ is replacing the observational treatment prediction $p(t \mid z_{1:n}, \mathscr{O})$ with the fully randomized assignment $p(t \mid \mathscr{E})$.

Application

We applied our methodology to study the effect of maternal smoking cessation during pregnancy on birthweight. Our dataset is an excerpt of the data studied in Almond (QJE 2005). The dataset contains information on singleton births in Pennsylvania between 1989 and 1991.

We define the treatment variable T to take the value 1 if the individual is assigned to the smoking cessation group, and 0 if they are instructed to continue smoking.

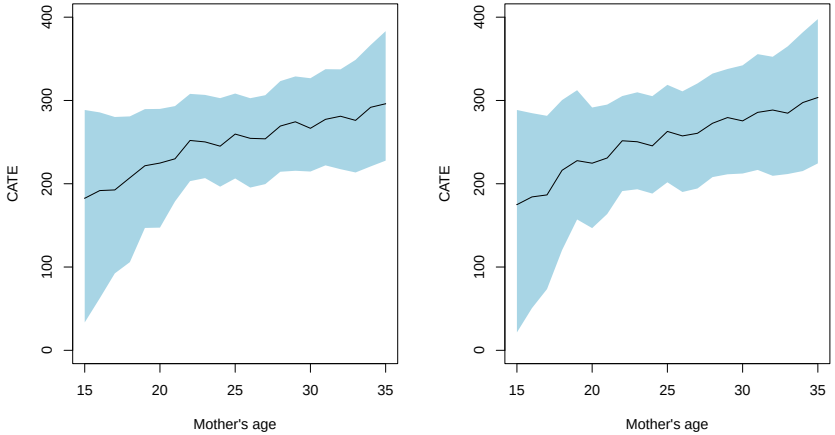


Figure: The posterior means and pointwise 95% intervals for the conditional average treatment effects given age (y-axis scale is in grams): BART and the Bayesian bootstrap (left); BART augmented with the clever covariate and the Bayesian bootstrap (right).

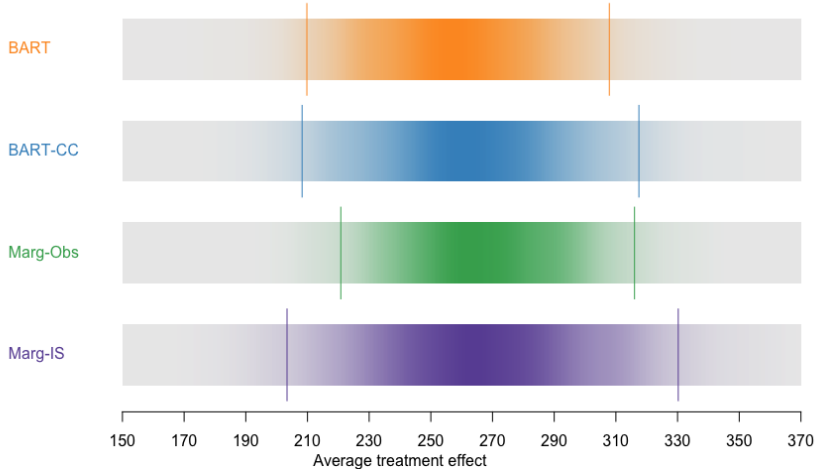


Figure: Posterior distributions for the average treatment effect (x-axis scale is in grams). The darkness of the strips is proportional to the posterior density, with the central 95% credible regions indicated.

Conclusions

- Target trial emulation (Hernan and Robins 2016) provides a powerful qualitative workflow for causal analysis of observational studies
- By augmenting the target trial protocol with a generative model we can simulate outcomes from the target trial, conditional on observational data, and pick off any (scientific) causal effects of interest
- The framework highlights that counterfactuals are not needed for causal inference on scientific (falsifiable) hypothesis – as a population-scale RCT is as good as it gets
- The usual causal assumptions map to interpretable and intuitive assumptions on transportability of predictive models across populations and conditions

References

Preprint: Yiu, A., Fong, E., Walker, S.G. and Holmes, C. *"Causal predictive inference and target trial emulation."* https://arxiv.org/abs/2207.12479

- Dagan, N. et al. (2021) BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *The New England Journal of Medicine,* 384:1412-1423
- Dawid, A.P. (2021) Decision-theoretic foundations for statistical causality. *Journal of Causal Inference,* 9:39–77
- Fong, E., Holmes, C. and Walker, S.G. Martingale posterior distributions (with discussion). *JRSS-Series B,* 2022.
- Hernán, M. A., and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8), 758-764.

Further reading: papers

- Rubin, D.R. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6(1), 34-58. The first rigorous formulation of the Rubin Causal Model. Frames causal inference as a mising data problem and advocates Bayesian inference for estimating causal effects.
- Holland, P.W. (1986). Statistics with Causal Inference. *JASA*, 86, 945-960. Overview of the RCM and the Fundamental Problem of Causal Inference.
- Dawid, A.P. (2000). Causal Inference without Counterfactuals (with discussion). *JASA*, 95, 407-424. Arguments against the use of counterfactuals for causal inference, with discussion from prominent advocates of counterfactuals.
- Observational Studies. Volume 8, Issue 2, 2022. Interviews with the most influential causal researchers: Judea Pearl, Don Rubin, James Heckmann, and Jamie Robins.
- Hernán, M.A., Wang, W. and Leaf, D.E. (2022). Target Trial Emulation: A Framework for Causal Inference From Observational Data *JAMA*, 328, 2446-2447. Recent overview of target trial emulation.

Joint modelling

We have advocated using a joint predictive distribution $p(z_{n+1:N} \mid z_{1:n})$ to predict the hypothetical target trial data. The emphasis here is to have dependence between the data points.

One might question the merit of having dependence in our predictive structure if the data are believed to be independent. The point is that we are (partially) using probability to quantify epistemic uncertainty; the dependence structure allows us to update our predictive as we acquire more data in order to iteratively improve our predictive accuracy for the future.

Exchangeability

Even if the data are believed to be i.i.d., the underlying data-generating mechanism is unknown, so we can model the data as being conditionally i.i.d. given an unknown parameter P. By averaging over a distribution on P (representing our epistemic uncertainty on P c.f. a prior in Bayesian inference), the resulting distribution on the variables is exchangeable, i.e.

$$Pr(X_1 = x_1, \ldots, X_k = x_k) = Pr(X_{\sigma(1)} = x_1, \ldots, X_{\sigma(k)} = x_k)$$

for all k and any permutation σ of $\{1, \ldots, k\}$. In words, the joint distribution is invariant under reorderings.

Under very general conditions, de Finetti's theorem tells us that the converse is also true; any exchangeable probability distribution can be derived from a conditionally i.i.d. model averaged over a mixing distribution on P.

What does it mean to be "Bayesian"

The target trial emulation approach uses predictive Bayesian models to make inference on causal parameters, but what does it mean to be Bayesian?

What does it mean to be "Bayesian"

I will define "Bayesian" as the use of joint probability for all statements of uncertainty – whether that be for parameters of interest or predictions on observables

There are many good reasons why priors are useful for inference – and we're **not** making any criticism of this. However, sometimes we may wish to avoid the prior specification step

We will explore whether one can be "Bayesian" without the specification of a prior, while acknowledging any implications of this

Predictive Inference

Our view is grounded in **predictive inference** that focuses on the joint predictive, $p(y)$, for observables as the primary tool for analysis

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta$$

This view involves working directly with the joint predictive, $p(y)$, to define statements of uncertainty on parameters of interest, θ, without necessarily going through a likelihood-prior construction

The approach has its roots in de Finetti, also Geisser, and others

- [Geisser, 1975, Roberts, 1965, Ericson, 1969, Geisser, 1982, Geisser, 1983, Hahn, 2015, Fortini and Petrone, 2020]

An illustration with traditional Bayes

A man walks into a bar (in Venice) and approaches the barman and asks him a question

The barman points over to a table where four Bayesians are deep in conversation

The man goes over to the table and says....

"Excuse me, the barman said you may be able to help. I found this unusual coin outside. I've tossed it 10 times and heads has come up 7 times. I'm trying to work out what is the chance that heads will come up on the next toss, and what uncertainty is there on the long running frequency of heads?"

Bayesian 1

The first Bayesian, (Prof O), says,

"Oh no, why oh why, did you tell me the outcome of the experiment! Now there's nothing I can do for you. Be gone."

Bayesian 2

The second Bayesian, (Prof L), says,

"Don't worry my friend, it's quite clear that you knew nothing about the properties of the coin before you found it and hence the probability of heads on the next toss, having seen 7 heads in 10 trials, is easily seen to be 2/3."

Bayesian 3

The third Bayesian, (Prof J), says,

"Don't listen to my colleague, he's a fool, because I wager you also knew nothing about the square root of the probability of heads for an unusual coin found in the street! Hence, the answer is straightforward, it is of course 0.68181818...."

Bayesian 4

The fourth Bayesian, (Prof G), says,

"Let's not get hung up on knowing the outcome of the experiment. Pull up a chair and let's have a discussion on your retrospective subjective beliefs on properties of coins that you happen upon in the street."

The answer is clear

At this point the man is seriously questioning his sanity....and is about to leave, when

A statistician in the corner (Prof F), smoking a pipe, says, don't listen to that rabble. You've seen 7 heads from 10 identical trials. Your best guess of the probability of heads on the next toss is 0.7. Your best prediction, in this situation, is the average of the outcomes that you've seen.

Thank you(!) says the man – that makes perfect sense

Happy Bayes

The Bayesians break out laughing, muttering about incoherence

They turn to themselves and spend the rest of the night drinking happily into the early hours of the morning and arguing about how to analyse a 2×2 contingency table of 4 numbers, without reaching consensus :)

Exchangeability and coherence

Of course, the traditional Bayesians are right in principle, as the coin tossing outcome is exchangeable

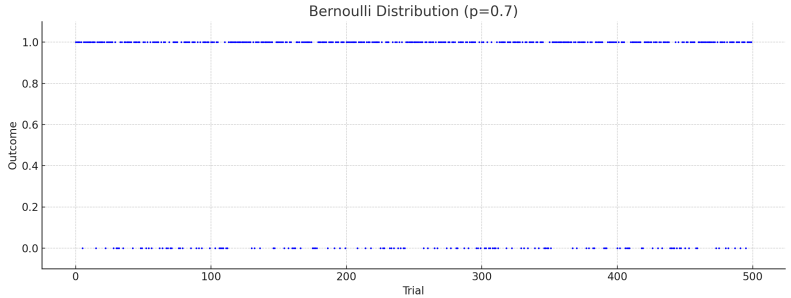


Figure: Underlying population of coin tosses extending to ∞

Our analysis method shouldn't, in principle, change depending on where the experiment of n tosses starts and stops

Posterior uncertainty

By following Bayes rule we have

$$p(\theta \mid y_{1:n}) \propto \prod_i p(y_i \mid \theta)\, p(\theta)$$

Questions on interpretation:

- What is the source of the uncertainty following the update?
- Consider again the coin tossing population

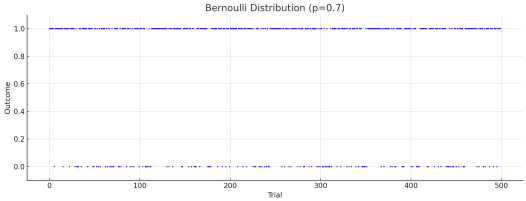


Figure: Underlying population of coin tosses extending to ∞

Statistical Uncertainty – a predictive view

All statistical uncertainty arises from data that you don't have

As, if you had all the data, if you'd measured every unit in the population, then you'd have the answer – no uncertainty

In which case, don't model the data you have – you already have it – model the data that you don't have that's needed to answer the question

Statistical Uncertainty – a predictive view

Make the data you don't have, from the remaining population, $y_{n+1:\infty}$, the focal point of the analysis

The uncertainty in any population parameter of interest, θ, flows from the missing $Y_{n+1:\infty}$

We will connect Bayesian reasoning with the "imputation" of the missing $Y_{n+1:\infty}$, and explore predictive (generative) models for inference

Bayesian inference directly targeting of the posterior

We can consider a data augmentation approach to expand the posterior distribution to **include data that we don't have** leading to the joint posterior

$$p(\theta, y_{n+1:\infty} \mid y_{1:n})$$

We write this conditional on the data that you have

We can think of

- $y_{obs} = y_{1:n}$ as the observed data
- $y_{mis} = y_{n+1:\infty}$ as the remaining (as if missing) data in the population
- $y_{comp} = \{y_{obs}, y_{mis}\}$ as the complete population

Remark: if we had y_{comp} then our inference problem would be solved

Predictive inference makes y_{mis} the focal point of the analysis

Predictive Inference – to obtain the posterior

Product rule,

$$\begin{aligned} p(\theta, y_{n+1:\infty} \mid y_{1:n}) &= p(\theta \mid y_{1:n}, y_{n+1:\infty})\, p(y_{n+1:\infty} \mid y_{1:n}) \\ &= p(\theta \mid y_{1:\infty})\, p(y_{n+1:\infty} \mid y_{1:n}) \end{aligned}$$

or if you prefer

$$p(\theta, y_{mis} \mid y_{obs}) = p(\theta \mid y_{comp})\, p(y_{mis} \mid y_{obs})$$

We can then obtain the posterior distribution of interest through marginalization (data augmentation)

$$\begin{aligned} p(\theta \mid y_{obs}) &= \int p(\theta, y_{mis} \mid y_{obs})\, dy_{mis} \\ &= \int p(\theta \mid y_{comp})\, p(y_{mis} \mid y_{obs})\, dy_{mis} \end{aligned}$$

Representation

The following equality holds (for any Bayesian posterior)

$$\begin{aligned} p(\theta \mid y_{obs}) &= \int p(\theta \mid y_{comp})\, p(y_{mis} \mid y_{obs})\, dy_{mis} \\ &= \frac{p(y_{obs} \mid \theta) p(\theta)}{p(y_{obs})} \end{aligned}$$

How to interpret the representation?

$$p(\theta, y_{mis} \mid y_{obs}) = p(\theta \mid y_{comp})\, p(y_{mis} \mid y_{obs})$$

We now have two components we need to define

- $p(\theta \mid y_{comp})$ – the posterior distribution arising if you had the total population data
- $p(y_{mis} \mid y_{obs})$ a ***joint*** predictive model for the unmeasured population given the data we have

Predictive Bayes

$p(\theta \mid y_{comp})$ contains no uncertainty as it conditions on the complete population

We can (loosely) write this as an empirical statistic, or relative frequency,

$$p(\theta \mid y_{comp}) \equiv \theta(y_{comp})$$

– this is the "lurking parameter" of de Finetti [de Finetti, 1937]

Predictive Inference

We can label the two components making the Bayes posterior as

$$p(\theta \mid y_{obs}) = \int \underbrace{\theta(y_{comp})}_{\text{Estimate (large sample)}} \underbrace{p(y_{mis} \mid y_{obs})}_{\text{Predictive (missing info.)}} dy_{mis}\,.$$

Bayesian inference is inherently predictive, as every posterior has this decomposition. In contract maximum likelihood estimation isn't predictive

This emphasises the source of statistical uncertainty in $p(\theta \mid y_{obs})$ as arising from the missing population measurements, $y_{n+1:\infty}$, needed to answer the inference question

We don't **have to** introduce a prior as the starting point is $p(y_{mis} \mid y_{obs})$

Constructing a Predictive

Consider the Bayesian joint predictive under a *prequential* factorization:

$$p(y_{n+1:\infty} \mid y_{1:n}) = \prod_{i=n+1}^{\infty} p(y_i \mid y_{1:i-1}).$$

From which we can impute $Y_{n+1:\infty} \sim p(y_{n+1:\infty} \mid y_{1:n})$ through the recursion

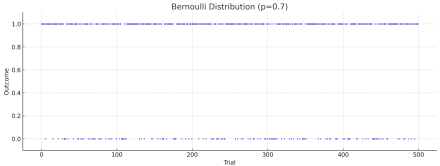
1. Draw $Y_{n+1} \sim p(y \mid y_{1:n})$
2. Draw $Y_{n+2} \sim p(y \mid y_{1:n}, Y_{n+1})$ from the updated predictive, ..., etc

This constructive specification for the joint requires a 1-step predictive $p(y_{n+1} \mid y_{1:n})$ AND the update $p(y_{n+2} \mid y_{1:n}, y_{n+1})$

Once we have $y_1, y_2, \ldots, y_N$ for large N then we can pick off the parameter of interest; and repeat

Comments on predictive resampling

The starting point of $p_n = p(y_{n+1} \mid y_{1:n})$ violates "coherent" belief updating as we use the data to choose the model



We can use $p_n = p_{\widehat{\theta}}(y_{n+1} \mid y_{1:n})$ where $\widehat{\theta}$ is an MLE [Holmes and Walker, 2023], or alternatively $p_n = \mathbb{P}_n$ the empirical distribution [Fong et al., 2023] – no priors in either case

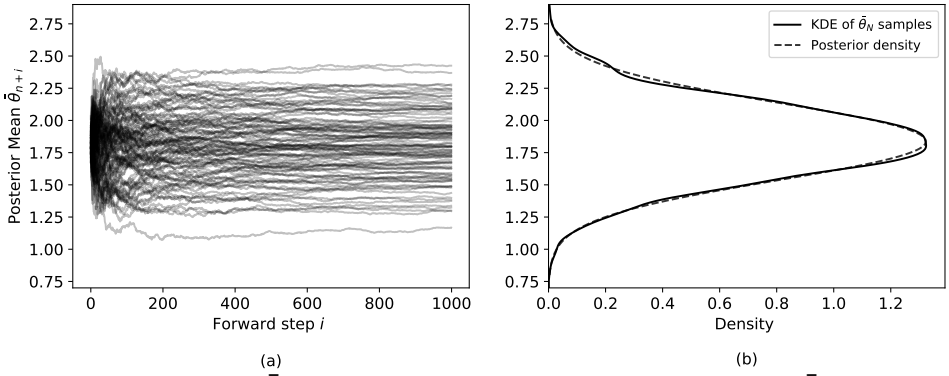
The update to the predictive p_i at each step necessitates the need for efficient online, continual, learning (model updating)

The algorithm is trivially parallel across samples of θ_N

A Parametric Example

Example

Let $p(y \mid \theta) = \mathcal{N}(y \mid \theta, 1)$, with $p(\theta) = \mathcal{N}(\theta \mid 0, 1)$.
Posterior predictive $p(y_{n+1} \mid y_{1:n})$ is normal, estimate is $\bar{\theta}_n = \sum_{i=1}^{n} y_i / n$.



Return of the man

The man returns to the bar the next day (the Bayesian's haven't left yet)

He seeks out Prof F, the pipe smoking statistician, and says

"Thank you so much for helping me yesterday. I found another unusual coin outside. I've tossed it 10 times and heads has come up 0 times. I'm trying to work out what is the chance that heads will come up on the next toss, and what uncertainty is there on the long running frequency of heads?"

Prof F looks perplexed, and rather angry,....

"go talk to the Bayesians"

Conclusions

The predictive viewpoint highlights $Y_{n+1:\infty}$ as the source of Bayesian uncertainty. Casts Bayesian inference as a missing data problem

Conclusions

The predictive viewpoint highlights $Y_{n+1:\infty}$ as the source of Bayesian uncertainty. Casts Bayesian inference as a missing data problem

The viewpoint is agnostic to the use of a prior. You're free to use a prior if it helps – whatever gives you the best conditional predictive

Conclusions

The predictive viewpoint highlights $Y_{n+1:\infty}$ as the source of Bayesian uncertainty. Casts Bayesian inference as a missing data problem

The viewpoint is agnostic to the use of a prior. You're free to use a prior if it helps – whatever gives you the best conditional predictive

Emphasizes the close link between Bayesian inference and ability to simulate observations from a system of interest (objectively)

Conclusions

The predictive viewpoint highlights $Y_{n+1:\infty}$ as the source of Bayesian uncertainty. Casts Bayesian inference as a missing data problem

The viewpoint is agnostic to the use of a prior. You're free to use a prior if it helps – whatever gives you the best conditional predictive

Emphasizes the close link between Bayesian inference and ability to simulate observations from a system of interest (objectively)

Places frequentist uncertainty, $\theta(Y_{1:n})$, and Bayesian uncertainty, $\theta(\{y_{1:n}, Y_{n+1:\infty}\})$, on an equal footing – making clear the essential distinction (which doesn't involve a prior)

Conclusions

The predictive viewpoint highlights $Y_{n+1:\infty}$ as the source of Bayesian uncertainty. Casts Bayesian inference as a missing data problem

The viewpoint is agnostic to the use of a prior. You're free to use a prior if it helps – whatever gives you the best conditional predictive

Emphasizes the close link between Bayesian inference and ability to simulate observations from a system of interest (objectively)

Places frequentist uncertainty, $\theta(Y_{1:n})$, and Bayesian uncertainty, $\theta(\{y_{1:n}, Y_{n+1:\infty}\})$, on an equal footing – making clear the essential distinction (which doesn't involve a prior)

E. Fong, C. Holmes, S.G. Walker, 'Martingale Posterior Distributions' JRSS-B Discussion paper, 2023

References (1)

de Finetti, B. (1937).
La prévision: ses lois logiques, ses sources subjectives.
In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.
[English translation in Studies in Subjective Probability (1980) (H. E. Kyburg and H. E. Smokler, eds.) 53-118. Krieger, Malabar, FL.].

Ericson, W. A. (1969).
Subjective Bayesian models in sampling finite populations.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 31(2):195–224.

Fong, E., Holmes, C., and Walker, S. G. (2023).
Martingale posterior distributions.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 85(5):1357–1391.

Fortini, S. and Petrone, S. (2020).
Quasi-Bayes properties of a procedure for sequential learning in mixture models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4):1087–1114.

References (2)

Geisser, S. (1975).
The predictive sample reuse method with applications.
Journal of the American Statistical Association, 70(350):320–328.

Geisser, S. (1982).
Aspects of the predictive and estimative approaches in the determination of probabilities.
Biometrics, pages 75–85.

Geisser, S. (1983).
On the prediction of observables: a selective update.
Technical report, University of Minnesota.

Hahn, P. R. (2015).
Predictivist Bayes density estimation.

Holmes, C. C. and Walker, S. G. (2023).
Statistical inference with exchangeability and martingales.
Philosophical Transactions of the Royal Society A, 381(2247):20220143.

References (3)


Roberts, H. V. (1965).
Probabilistic prediction.
Journal of the American Statistical Association, 60(309):50–62.